DOCUMENT RESUME

ED 361 377 TM 020 472

TITLE	Educational Testing: The Canadian Experience with Standards, Examinations, and Assessments. Report to Congressional Requesters.
INSTITUTION	General Accounting Office, Washington, DC. Program Evaluation and Methodology Div.
REPORT NO	GAO/PEMD-93-11
PUB DATE	Apr 93
NOTE	77p.
AVAILABLE FROM	U.S. General Accounting Office, P.O. Box 6015, Gaithersburg, MD 20884-6015 (first copy is free; additional copies are $2; orders for 100 or more to a single address are discounted 25 percent).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Academic Achievement; *Academic Standards; *Educational Assessment; Educational Testing; Elementary School Students; Elementary Secondary Education; Financial Support; Foreign Countries; National Competency Tests; National Programs; Public Opinion; Public Support; Secondary School Students; *Student Evaluation; Test Construction; *Testing Programs; Test Results; Test Use
IDENTIFIERS	Canada; *Large Scale Programs; *Provincial Examinations; Standard Setting; United States

ABSTRACT

The experience of the Canadian provinces with large-scale testing programs is reviewed to suggest answers to questions currently discussed in the United States regarding the proposed national system of achievement testing. In the Canadian provinces, tests are linked to provincial curricula. Standards and assessments are not set by experts with no direct responsibility for implementing curriculum and instruction, but are prepared with the cooperation of teachers and administrators. Assessments cover broad subject areas and monitor the overall education system, and examinations certify individuals' mastery of specific high school courses. The provinces have not found it necessary to attach high stakes to all tests. Examination results are used in conjunction with teacher-assigned grades, and assessments have no consequences for individual students. Safeguards have been developed for each type of test; and provincial school funding formulas, which are independent of the testing process, tend to level resources among schools within a province. Most Canadian educators and the public support the testing programs, but they have no hard evidence that the programs lead directly to improved instruction. Appendixes list characteristics of the provincial programs, participants in site and telephone interviews, and major contributors to the report. Fourteen tables provide information about the testing programs. (Contains 8 references.) (SLD)

United States General Accounting Office

GAO

Report to Congressional Requesters

April 1993

ED 361 377

EDUCATIONAL TESTING

The Canadian Experience With Standards, Examinations, and Assessments

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM020472

GAO/PEMD-93-11

BEST COPY AVAILABLE

TM

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-249070

April 28, 1993

The Honorable William D. Ford
Chairman, Committee on Education and Labor
House of Representatives

The Honorable William F. Goodling
Ranking Minority Member, Committee on Education and Labor
House of Representatives

The Honorable Dale E. Kildee
Chairman, Subcommittee on Elementary, Secondary,
and Vocational Education
Committee on Education and Labor
House of Representatives

Recently, various individuals and organizations, including the National Council on Education Standards and Testing, have proposed a national system of achievement testing as the cornerstone of an overall approach to improving education. At your request, we reviewed the experience of the Canadian provinces with large-scale testing programs to suggest answers to questions discussed currently in the United States regarding this proposal. In the Canadian provinces, tests are linked to provincial curricula and measure the extent to which students have achieved provincial standards developed through the widespread involvement of teachers, provincial officials, and subject-area experts. Canadians have developed a system of tests, stakes, and safeguards designed not only to emphasize the importance of the tests but also to protect students from faulty testing and the misuse of test results.

We are sending copies of this report to officials at the Department of Education and to others who are interested, and we will make copies available to others upon request. If you have any questions or would like additional information, please call me at (202) 512-2900 or Robert L. York, Director of Program Evaluation in Human Services Areas, at (202) 512-5885. Other major contributors to this report are listed in appendix III.

Eleanor Chelimsky

Eleanor Chelimsky
Assistant Comptroller General

Executive Summary

Purpose

In its final report to the Congress in 1992, Raising Standards for American Education, the National Council on Education Standards and Testing (NCEST) joined other groups in proposing a national system of standards and assessments as the cornerstone of an overall approach to improving precollege education. NCEST maintained that this new system, consisting of clusters of states using improved testing methods to measure progress against the standards, has the potential to improve education by raising expectations for what should be learned and by increasing the accountability of students and educators.

At the request of the House Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education, GAO reviewed Canada's experience with large-scale testing programs. In this report, GAO examines the Canadian experience in terms of five questions about testing policy that are currently being discussed in the United States. These are: (1) How have educational standards been set, by whom, and at what level? (2) What kinds of tests have been used to assess whether the standards are being met? (3) What types of stakes have been attached to tests to ensure that they will be taken seriously? (4) What explicit safeguards have been used to prevent misuse of tests? (5) Have efforts at raising expectations and checking results brought promise of improved teaching and learning?

Background

The last two decades have witnessed both growing concern about the quality of U.S. education and a variety of school improvement efforts. Currently, attention has focused on national standards and a system of related assessments as key components of education reform. During this same period, improvement efforts in many Canadian provinces have also been directed toward developing standards and expanding assessment and examination programs that resemble those being discussed for the United States. GAO's study of the Canadian experience suggests both implementation issues and possible results the United States may encounter in adopting similar policies and methods. Information for this study was obtained through interviews with provincial education officials, researchers, and members of the business community and through analysis and review of provincial documents and evaluation reports.

Results in Brief

In Canada, standards are set at the provincial rather than at the national level. Tests are typically tied to each provincial curriculum and measure the degree to which students have achieved specified provincial standards

set by teachers, subject-area experts, and provincial education officials. Unlike current practice in the United States, standards and assessments are not established by groups of experts who have no direct responsibility for implementing curriculum and instruction.

Provinces use two different types of tests for two different purposes: assessments cover broad subject areas and monitor the overall education system, and examinations certify individuals' mastery of specific high school courses. Each has characteristics suited to its specific purpose. The Canadian provinces have not found it necessary to attach high stakes to all tests. In the case of high school examinations, scores are used only when combined with teacher-assigned grades to determine final grades and, thus, help determine student placement, grade promotion, and postsecondary opportunities. The provincewide assessments have no consequences for individual students and are used to monitor the system.

Safeguards have been developed for each type of test to protect individuals from unfair testing practices and misuse of test scores. Provincial funding formulas, although independent of the testing programs, tend to level resources among schools within a province. Thus, in contrast to the United States, Canadian practices prevent the gross disparities in resources among districts that raise concerns regarding the equity of students' opportunities to learn the materials tested. Although most Canadian educators and the public support testing programs, they have no hard evidence nor independent yardstick demonstrating that the tests themselves have directly improved instruction or learning.

Principal Findings

Curriculum Standards Reflected by Tests

In most provinces, teachers, test specialists, and subject-area experts, under the direction of provincial education officials, review the curriculum to develop test specifications that reflect the relative importance of different learning objectives in a course of study. Widespread teacher involvement is also common in both writing test questions and serving on central grading panels. This involvement helps increase teachers' knowledge of curricula and instruction and aids in the development of tests that are compatible with good classroom instruction. In contrast, U.S. teachers do not typically play key roles in the development of commercial or state tests and, thus, do not have access to similar experiences that hold the promise of improving both teaching and testing.

In most Canadian provinces, new tests are developed for each test administration, whereas in the United States, the same tests are commonly used for several years. Using the same tests increases the predictability of what will be tested and, consequently, the likelihood that teachers will focus their instruction on particular test items rather than on the content or skills that the items represent. This practice has been associated with decreasing the usefulness and the generalizability of what is learned as well as the value of the test for measuring content or achievement of skills.

Only after much experience with developing standards at the provincial level has Canada identified a set of objectives that can be applied to measure achievement in all provinces and developed an instrument for use Canada-wide. In both the Canada-wide and provincial test development work, a contrast is already observable between U.S. and Canadian methods in this area: Canadian tests are typically produced to reflect established curricula and for a specific purpose. U.S. tests are not usually tied to established standards.

Different Tests for Different Purposes

Eight provinces use assessments to monitor achievement in broad subject areas at selected grades, usually at the elementary and middle school level. Six provinces use sampling strategies in administering all or parts of these tests. Assessments typically cover one subject area for each grade level targeted for the assessments. This permits, for the same cost, covering more content in greater depth than could be done by testing all students in many subjects in a given year. Diploma examinations are used by five provinces to certify the achievement of individual high school students in specific academic courses and are therefore administered to all students enrolled in the courses.

Low- and Limited-Stakes Testing

The Canadian experience indicates that high stakes are only one way of reinforcing the importance of the tests and what they measure. In fact, the most sizable testing efforts in Canada are the low-stakes assessment programs. The widespread involvement of teachers in developing tests and in centralized grading activities appears to have increased the acceptance of the tests and their influence upon instruction. In the case of the high-stake diploma examinations, three of the five provinces have reduced the contribution of the examination to the course grade to reflect the importance of classroom work.

The Presence of Safeguards

Canadians have developed safeguards to help prevent the misuse of test results. For tests that monitor the education system, safeguards, such as aggregated reporting of results and delayed feedback, are designed to prevent the misuse of scores as evaluations of individual teachers and students. In addition, the low stakes attached to these tests reduce the likelihood that instruction in content areas targeted for assessment will be unduly emphasized at the expense of equally important content that is not. For tests that certify students, safeguards are designed to protect them from arbitrary test practices, ensure multiple opportunities for success, and accommodate those with disabilities.

Testing and Improved Learning

At this time, the Canadian provinces lack an independent yardstick with which to monitor changes in achievement over time. Until such a measure is in place, there are no data that might reveal a relationship between increased testing and improved learning. However, provincial educational agencies have achieved greater control of classroom instruction as teachers have aligned their teaching to the tests. Whether this uniformity results in improved learning remains subject to different opinions. Provincial education agency officials and teachers agree that teachers' involvement in the testing programs has helped increase their knowledge of the curriculum and has improved the tests.

Despite the lack of evidence for a direct association between testing and achievement, the establishment of new testing programs has been a popular response to perceived educational declines. In the 1970s, only 2 of the 10 provinces offered provincial examinations. Currently five provinces have examination systems and a sixth province plans to develop a system. The number of provinces with assessments in place increased from five in 1980 to eight in 1992, and an additional province plans to initiate a program in 1993.

Recommendations

This report contains no recommendations.

Agency Comments

Officials in several Canadian provincial departments and other experts on Canadian educational testing and assessment reviewed a draft of this report. Their comments were incorporated where appropriate. These reviewers generally agreed with GAO findings.

Contents

Executive Summary		2
Chapter 1 Introduction		8
	Objectives, Scope, and Methodology	8
	Relevance of the Analogy	9
	Canadian Schools and Testing	11
	Organization of the Report	14
Chapter 2 Linking What Is Taught and What Is Tested		15
	Testing in Alberta	17
	Testing in British Columbia	22
	Summary and Conclusions	24
Chapter 3 Stakes and Safeguards Associated With Canadian Tests		26
	Examination Stakes	27
	Safeguards for Students in Design and Scoring of Examinations	30
	Assessment Stakes and Safeguards	32
	Summary and Conclusions	35
Chapter 4 The Promise of Testing for Improving Teaching and Learning		37
	Observed Changes in Teaching	37
	Observed Changes in Student Learning	38
	Support for the Examinations	39
	Summary and Conclusions	40
Chapter 5 Steps Toward the Development of a Canada-Wide Assessment of Literacy and Numeracy		42
	Origins and Purposes of the National Indicators Project	42
	Efforts to Structure the Project	43
	Getting Started in Test Development—Standards	44
	Development of the Assessment Instrument	45
	Reporting Results	46
	Feasibility Issues	46
	Summary and Conclusions	47

Chapter 6 Conclusions

49
Our Research Questions About National Tests 49
Canada's Experience 49

Appendixes

Appendix I: Characteristics of Canadian Provincewide Testing Programs 54
Appendix II: Expert Participants in Site and Telephone Interviews 70
Appendix III: Major Contributors to This Report 73
Bibliography 74

Tables

Table I.1: Alberta Diploma Examination System 55
Table I.2: Alberta Assessment Program 56
Table I.3: British Columbia Provincial Examination System 57
Table I.4: British Columbia Provincial Learning Assessment Program 58
Table I.5: Manitoba Provincial Examinations 59
Table I.6: Manitoba Curriculum Assessment Program 60
Table I.7: New Brunswick Provincial Achievement Examination 61
Table I.8: New Brunswick Oral Proficiency Interview for the Second Language Project 61
Table I.9: Newfoundland and Labrador Public Examinations 63
Table I.10: Newfoundland and Labrador Achievement Testing Program 64
Table I.11: Nova Scotia Achievement Tests 65
Table I.12: Ontario Student Assessments 66
Table I.13: Quebec Ministry-Prepared Examinations 68
Table I.14: Quebec Curriculum Tests 69

Abbreviations

CMEC	Council of Ministers of Education, Canada
GDP	Gross domestic product
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCEST	National Council on Education Standards and Testing
NCTM	National Council of Teachers of Mathematics
TIMSS	Third International Mathematics and Science Studies

Chapter 1

Introduction

In recent years, a growing number of educational and political leaders have come to advocate some form of national student achievement testing for the United States, and many also call for improved testing methods and coverage of more rigorous content as key parts of an overall approach to improving education. Beyond testing, other related proposals include establishing national standards for what is to be taught and learned and, less often, raising the importance of tests by the eventual broader use of test scores for decisions regarding promotion, graduation, postsecondary education, and employment opportunities. Together, this set of reforms is believed to form a potent combination of pushes and pulls that holds a promise of improving the quality of teaching and learning in U.S. schools. Proposals along these lines have come from, among others, the congressionally mandated National Council on Education Standards and Testing (NCEST).[1]

Objectives, Scope, and Methodology

The House Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education asked us to examine several broad topics regarding national tests of school achievement. In another report, we used surveys of states and school districts to describe the extent and cost of present testing in the United States, the possible overlap of new tests with old, and other effects.[2] Still another report examines the validity and reliability of the current use of U.S. test data to assess student progress toward nationally defined achievement levels.[3]

In the present report, we deal with the last part of the Committee's request, regarding the lessons to be learned from the experience of other nations that have implemented large-scale testing efforts. From our knowledge of the discussions of proposals for national standards and testing in the United States, we identified five specific matters at issue, and we organized our review around them:

- How have educational standards been set, by whom, and at what level?
- What kinds of tests have been used to assess whether the standards are being met?

[1]The Congress established the Council in Public Law 102-62 to report on the feasibility and advisability of national tests of school achievement. The Council did its work in 1991 and submitted its conclusions in Raising Standards for American Education, A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People (Washington, D.C.: January 1992).

[2]Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination (GAO/PEMD-93-8, January 13, 1993).

[3]We reported interim findings in National Assessment Technical Quality (GAO/PEMD-92-22R, March 11, 1992). The full report of our work is forthcoming.

- What types of stakes have been attached to tests to ensure that they will be taken seriously?
- What explicit safeguards have been used to prevent misuse of test results?
- Have efforts at raising expectations and checking results brought promise of improving teaching and learning?

By agreement with the Committee, we decided to limit our search for useful international comparisons on these matters to Canada alone, in part because, as we discuss below, the Canadian experience is perhaps more relevant for the United States than that of any other country, but also because of resource constraints.

To understand the Canadian experience, we gathered firsthand data in Canada, conducted interviews in person and by telephone, and examined a wide range of published and unpublished documents and literature. Specifically, we visited three provinces (Alberta, British Columbia, and Ontario), and in each one, we interviewed officials of the provincial departments of education, university scholars, testing experts, teachers, business leaders, and others.[4] We interviewed individuals in the remaining seven provinces by telephone or in Washington, D.C. We reviewed relevant provincial education department documents as well as published research and evaluation evidence. For comparative data on issues in testing in the United States, we drew on information gathered in developing our other reports cited previously, and we conducted additional interviews with officials in 10 states on the topic of teacher involvement in standards and testing projects. Officials in several Canadian provincial departments and other experts on Canadian educational testing and assessment reviewed a draft of this report. We incorporated their comments and made corrections when appropriate. These reviewers generally agreed with our findings concerning the Canadian experience.

Relevance of the Analogy

When very significant policy changes are considered and analysis of their feasibility or possible effects is needed, if the proposals are dramatically different from current practice there may be no relevant experience within the nation to guide such assessment. In the health care area, for example, many observers (including us) are examining major policy alternatives used in other nations to deal with costs and quality of care. Just such considerations guided our examination of the Canadian experience with educational testing.

[4]Canadian provinces use different terms for their education agencies. Six are departments, and four are ministries. We use the term department in this report when referring to both.

As an affluent, high-tech industrial society, Canada closely resembles the United States in per capita output, market-oriented economic system, and pattern of production involving a skilled labor force, and both countries thus face continuing concern for the effectiveness of their education systems.

Both nations' constitutions decentralize education decisions to the state or province. The scale of the schooling enterprise is of course very different—5 million Canadian students as compared with 41 million U.S. students—and this may permit Canada to spend more, proportionally, on education than we do—6.7 percent of gross domestic product (GDP) in 1986 compared to 4.8 percent in the United States—although Canadians have been willing to allow public expenditure of all kinds to grow to a higher level as a percent of GDP than we have (47 percent compared to 37 percent). Urban education concerns may appear less pressing in Canada, as only three cities have over a million people. Yet, just as in the United States, Canadian society struggles with political and cultural concerns arising from the diversity of its population including the 27 percent of French origin, the 11 percent Asian, Hispanic, and African-Canadians, and the indigenous groups that make up 1.5 percent.

In addition to these similarities between the nations, the testing policies and methods in place in many Canadian provinces have features that mirror those included in testing proposals now suggested for the United States. NCEST, for example, advocates not a single national test but a decentralized system of tests linked to what is taught in a state or region that would (1) measure the progress of students in reaching educational standards, (2) use test methods other than multiple-choice questions, and (3) involve teachers much more than is now common in all phases of testing. The Canadian testing experience provides examples of all of these features.

The United States, of course, does not lack experience with widespread testing programs of various types. In fact, several states have developed or are developing testing programs along the lines recommended by NCEST, which resemble to some degree tests used in the Canadian provinces. However, the testing programs here and in Canada contrast in terms of key characteristics—the extent of linkage of tests to a prescribed curriculum, the scope and prevalence of the test programs, and the degree of teacher involvement. As we looked at specific Canadian testing programs, we judged them to be examples of directions suggested for the United States and thus likely to afford contrasts on important

characteristics from which we could learn. (General background on Canadian education and testing are discussed in the section that follows.) These examples include the following contrasts:

- Canada and the United States are alike in using standardized testing at the state or province level. Most tests in the United States, however, are commercially developed tests of broad subject knowledge that have been only modestly customized to reflect what is taught in a state. Typically, these tests consist of only multiple-choice questions. Canadian provincial tests are developed within each province to be much more closely linked to a province's standards. Diverse testing methods are used.
- Canada and the United States are alike in supporting the development of new testing instruments. Most tests in the United States are in use for several years, whereas in the Canadian provinces, most tests are redeveloped for each administration. (New York is notable as one of the few states that develops a very large number of tests on an annual cycle.[5])
- Canada and the United States are alike in using some high-stakes tests, typically in the later years of high school. Most such tests in the United States, however, are multiple choice and many cover only minimal competencies. (New York is again notable as one of the few states that test advanced high school course knowledge using complex exam formats.) Provincial examinations differ by using multiple methods to test students' knowledge of diverse high school courses.
- Canada and the United States are alike in involving teachers in province and state testing work. The extent of their role, however, is different. Specific states (and testing firms) may involve teachers every few years as test specifications and questions are being developed. (New York involves a very large number of teachers on an annual cycle.) Canadian practices, in many provinces, differ markedly by relying on large numbers of teachers annually, at various stages of test development.

Thus, the Canadian testing experience, which differs from that in the United States although the settings are somewhat similar, can suggest both implementation issues and possible results the United States may encounter in adopting similar testing policies and methods.

Canadian Schools and Testing

When the four original provinces of Canada (New Brunswick, Nova Scotia, Ontario, and Quebec) were united in 1867, the individual provincial legislatures were given exclusive constitutional jurisdiction over

[5]In the case of the New York Regents Examinations, three exams are developed for each course specified for examination. A new form is therefore available for administration in the fall, spring, and summer semesters.

education. Each province has a ministry or department of education, which is headed by an elected member of the provincial legislature. Unlike the United States and most other nations, Canada has no national-level government agency responsible for education.

Responsibilities for elementary and secondary education are divided between the provincial governments and local school boards. The provincial departments set educational policy and standards and ensure adequate educational opportunities. Thus the departments certify teachers, develop curricula, and do research. The responsibilities of the school boards, which are delegated by the provincial legislatures or education departments, are somewhat similar to those of U.S. boards, including daily management of the schools, hiring staff, negotiating salary schedules, providing student transportation, and building and maintaining schools. More school funds are raised and distributed provincewide than in the states, so resource inequities within a province are much less marked than within some states where local tax revenues differ widely and form a significant part of school funding.

Education in Canada shares many other structural characteristics with that in the United States. School attendance is compulsory for about 10 years—children begin school at either age 6 or 7 and must remain in school through age 15 or 16. In both countries, grades are typically organized into elementary, middle, and high schools, and the school year is relatively short, 178 days on average in the United States and 188 days on average in Canada (although the typical U.S. school day averages about 30 minutes longer).

Provincial testing programs, under various names and in numerous forms, have existed in Canada since the first public high schools. Examinations were used originally to identify students who were considered academically eligible for high school. Later, the examinations became tests of school leaving and were used to determine eligibility for high school graduation. Eventually, the examinations were used to determine graduation status, qualify students for university admission, assess the effectiveness of high schools, and provide accountability to the public for the expenditure of taxpayer funds.

All provinces except Quebec and Newfoundland discontinued provincial examinations in the 1970s. This break with convention was linked to several factors. First, in Canada as well as elsewhere, the late 1960s and early 1970s were characterized by a movement toward political

decentralization and increased local control. Second, public confidence in Canadian schools grew as school district consolidation created an economy of scale which could support district-level specialists with expertise for oversight and guidance of teaching and learning. Third, more rigorous teacher certification requirements and better teacher pay led to a more educated teaching corps. Finally, educators escalated their criticism of the examination systems, questioning the validity of the examinations for the multiple purposes they were serving and claiming the examinations were damaging to instruction and learning.

However, in response to public concerns about the erosion of educational standards and demands for increased accountability, Alberta and British Columbia reintroduced examinations in the 1980s.[6] Manitoba introduced examinations in 1991. (Two others—Newfoundland and Quebec—retained their exams through the 1970s, the period when exam programs were dropped by most provinces.) New Brunswick is currently considering plans for a new system. Canada's experience with this burst of testing activity is the focus of our review.

As in many states, Canadian provinces have, in addition to the examinations, a second sort of test with lower stakes and for different purposes. These assessments or achievement tests have been in place in some provinces since the 1940s. In one province (Newfoundland), the assessments that are taken by the largest number of students resemble the commercial, norm-referenced, multiple-choice tests used in the United States. Other assessments are more like the provincial examinations in that they are developed by teachers, are redesigned for each administration, measure provincial curriculum standards, and they include written-response items. Still others combine features of both models. Despite their different forms, these tests share a common purpose—to assess the adequacy of the curriculum and instruction at the provincial level. Unlike the examinations, these tests are not intended for use as certifying or placement instruments. The assessment programs rose in importance as many examination programs were dismantled and concerns for accountability increased. Currently, 8 of the 10 provinces have an assessment program and a ninth (Saskatchewan) is developing a program for 1993. Only Prince Edward Island has no plans to reintroduce a provincewide assessment program.

[6]The Ontario Ministry of Education announced a return to a graduation examination system before similar decisions were made in Alberta and British Columbia. The plan was dropped in the face of teachers' opposition.

Organization of the Report

Chapter 2 describes provincial standards and the two major types of testing that address the standards. Chapter 3 discusses the consequences or stakes that are attached to different tests and the safeguards Canadian officials have designed to prevent misuse of results. Evidence on any effects of expanded testing in improving teaching and learning is presented in chapter 4. In chapter 5, we discuss the development of a new type of assessment for use Canada-wide. Finally, chapter 6 summarizes our study findings and draws implications for the U.S. discussion of national testing.

Chapter 2

Linking What Is Taught and What Is Tested

In this chapter, we answer evaluation questions 1 and 2: How have educational standards have been set, by whom, and at what level; and what kinds of tests are used to assess whether the standards are being met? In Canada, the provincial education departments are responsible for establishing the curriculum with provincial officials, teachers, and subject-area specialists sharing different responsibilities from province to province. At least some tests are tied to the curriculum in all eight provinces where provincial testing programs are in place; in six of them, all provincial tests are directly linked to the curriculum standards.[1]

Thus, a crucial distinction between Canadian provincial testing programs and most testing in the United States is that most provincial tests are designed to measure whether students learned what a specific curriculum directed they be taught, not whether students have mastered a general body of knowledge (as is common in the United States). The intentions behind such curriculum-linked tests are clear: teachers may be more likely to cover the intended curriculum when their students are to be examined on it. In addition, the tests themselves serve as vivid examples of what knowledge and skills should be taught and at what level this content should be assessed. Drawing on the Alberta and British Columbia testing programs as examples, this chapter emphasizes two main issues with regard to our evaluation questions: the linkage between curriculum and tests and the significant part played by teachers at various stages in test development throughout Canada.

Another crucial distinction between provincial test programs and most state test programs is that in most provinces, new tests are developed for each administration. In the United States, the same instruments are commonly used for several years. Thus, in Canada, the use of new tests may be expected to reduce the predictability of what will be tested and, therefore, the likelihood that instruction will emphasize answers to specific test items rather than focus upon the content or skills that the items were developed to represent. In the United States, however, the predictability of the test content may lead to the teaching of specific items. Current research has demonstrated that the alignment of instruction to particular test items will improve students' performance on those items; however, similar gains are not evident when other measures are used to assess the same content or skills. This suggests that achievement measured by test items that were the focus of instruction is not likely to

[1]Appendix I includes more information on all 10 provinces' testing programs, including type of tests used, courses and grades targeted for testing, types of safeguards and scoring procedures, and how each effort evolved over time.

carry over to other activities and that gains noted on these tests may well be illusionary.[2]

Provinces differ in their testing efforts in who does what (the roles played by teachers, the provincial education department contractors, and the public), in the tests themselves, and in choices made involving trade-offs (for example, use of the tests at the individual versus the provincial level, local validity versus broader comparability, or technical quality versus cost).

Both Alberta and British Columbia have two types of test programs commonly found in other provinces, an examination program and an assessment program. The first tests individual high school students' achievement in particular courses to certify individual achievement, and the other tests the performance of students throughout the province in subject areas at various grade levels as part of a curriculum evaluation process.

Half the provinces (5 of 10) have examination programs where we noted the following common features:

- The examinations are based on the curriculum, and both the content covered and the performance standards expected of students are determined by the provincial education department and subject-area teachers, with different provinces using various types and amounts of assistance from external groups.
- New examinations are developed each year, and different forms are prepared for each semester that a targeted course is offered. This has three purposes: to maintain test security, to encourage the responsiveness of examinations to changes in curriculum priorities, and to discourage teaching to the test.
- All examinations contain a mix of multiple-choice and written-response items.
- Examination grades determine between 30 percent and 50 percent of high school students' final grades in courses for which examinations are required and therefore contribute substantially to students' graduation status.[3]

[2]Koretz et al., "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests." Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education (Chicago: April 5, 1991).

[3]These provincial high school examinations are called diploma examinations in Alberta, provincial examinations in British Columbia and Manitoba, public examinations in Newfoundland, and ministry-prepared examinations in Quebec.

Most provinces (8 of 10) also have an assessment program designed to monitor learning and teaching throughout the province. We noted greater variation in these efforts; for example:

- Six of the eight provinces use curriculum-based, criterion-referenced tests; another uses a commercially developed test that is not directly linked to the provincial curricula; and yet another uses a test that consists of both curriculum-based and non-curriculum-based items.
- Provinces vary in how long they will use a particular assessment instrument, from only once to several years.
- Testing methods vary, with some tests consisting of mixtures of open-ended and multiple-choice items and others of only one type.
- Testing schedules vary, with some provinces testing according to prescribed schedules and others testing according to perceived need and budgetary constraints.
- Test coverage varies, with some provinces assessing achievement in depth in one subject area at a time and others examining achievement in several subject areas in a less comprehensive fashion.

Our analysis concentrates on approaches developed by Alberta and British Columbia for several reasons. First, both provinces have examination and assessment programs with technical and policy features like those being discussed in the United States.

- Tests are linked to standards, both for content covered and student performance.
- Tests include diverse measurement approaches beyond multiple-choice questions.
- Some of the tests carry high stakes for students.

Second, the programs in these provinces are comprehensive in that several grades and all major subject areas are targeted for testing on regular cycles. Finally, practices developed by these provinces are highly regarded and exert continuing influence on practices elsewhere in Canada.

Testing in Alberta

Alberta is Canada's fourth most populous province, with a population of about 2.6 million.[4] Approximately 80 percent of the population lives in urban areas. All grade 12 students are required to take at least one examination in either a college preparatory or general language arts

[4]The K-12 school enrollment of about 400,000 in Alberta is comparable to that of Utah.

course to graduate. Students enrolled in other examination courses (biology, chemistry, physics, French language arts for French-speaking students, mathematics, and social studies) are also required to take the examinations.[5] Translations of the examinations are available in French for French-speaking students or students who wish credentials in French. With regard to assessments, these are given in English, mathematics, science, and social studies to students in grades 3, 6, and 9.[6] Subject areas are tested on a 4-year cycle; each year, one subject area is targeted for assessment, and a new assessment is developed for that purpose. All students in targeted grades complete the multiple-choice component of the assessment; a sample of students in each grade completes a set of open-ended, performance-based items.

The Alberta Examination System

The Alberta Department of Education plays the key role in deciding what to test and how, but it gets a good deal of help. For each examination course, a panel of provincial education department staff and "seconded" teachers (hired from school districts to work for the department for 1-2 years) reviews the provincial course curriculum to draw up examination specifications, which address the following questions:

- What knowledge and skills should students have after the course is completed?
- What level of difficulty should be required of students?
- What type of items (multiple choice, short answer, or extended written response) should be used to measure achievement?
- How much of each part of the test will contribute to students' overall grades for a course?

In contrast to U.S. practice, where test development is often done by experts outside the public school system, teachers and specialists in Alberta school districts review the panels' exam specifications to build consensus that they are valid.[7] The specifications are also reviewed by representatives of different branches of the provincial education department, who can also make additions or changes. Approval by this

[5]The number of students enrolled in courses for which examinations are required varies. For example, about 24,300 students were required to write the academic English examination, while about 8,600 wrote the physics examination. The average number of tests taken by students enrolled in academic English is about 3.6; the average for students taking general English is about 1.6.

[6]Students in French immersion programs and French schools are assessed, at two grade levels, in French.

[7]The process includes obtaining comments from subject-area teachers, administrators, and school board members; formal reviews by teacher committees; and a second distribution of revised specifications to the schools for further comments from staff and board members.

group is required before the specifications are submitted for external review.

During the external review, the specifications are studied yet again, this time by a committee composed of representatives from organized educational stakeholder groups, including the provincial federation of teachers, the school supervisors' council, public colleges and universities in Alberta, and the Alberta Education Department.[8] This committee has the authority to recommend additional revisions, if necessary, and then approves the final specifications. The approved specifications are distributed to teachers of examination courses and students in these courses. Thus, the examinations are—at least potentially—closely linked to teaching and learning.

Test development specialists (employees of the provincial education department or secondees) next chair committees of classroom teachers hired to write items based on the specifications.[9] The specialists oversee the writing, edit the items, and assemble pilot forms of the tests. Using results from the field tests, a formal item analysis is prepared to determine the level of difficulty and to identify items that may be biased.[10] The specialists then use the pretest analysis to prepare the final forms of the examinations.[11] The examinations are submitted to the provincial education department and subsequently to the external review committee for final approval.

[8]Future plans call for involving the public by convening committees of parents and professionals working in fields related to the examinations to advise on the exam specifications.

[9]Selection for this work is rigorous. Those selected must hold a permanent teaching certificate, be teaching the exam course at that time, and have at least 2 years' experience doing so. Teachers apply to local superintendents, who nominate candidates to the provincial education department. The department selects teachers with leadership skills who represent the general demographic characteristics of the provincial population. A balance is sought between teachers who have had previous item-writing experience and those who have not.

The provincial education department pays the selected teachers U.S.$15 per hour plus expenses for work done during nonclass hours. For work done during the school day, the department pays the teacher a U.S.$51-per-day honorarium plus expenses and reimburses the school district for the cost of hiring a substitute for that teacher.

[10]About 250 students participate in each field test, and some examinations, particularly those in language arts, are field-tested three or four times because of difficulties in writing items that require written interpretation. Students enrolled in courses requiring examinations participate in the field tests of examinations that will be used the following year.

[11]A different exam is used each semester, for security. Three alternative forms of each examination are usually prepared, to accommodate the standard pattern in many Alberta schools of teaching courses for one semester only, rather than for a whole year as in the United States. The third version is available for summer students.

Teachers are also involved in confirming cutoff points set to determine standards of excellence (80 percent) and passing (50 percent) and in grading open-ended questions on the exams, serving in groups assembled just for this purpose. Generally three independent readers score each set of open-ended questions.[12] Teachers involved in grading determine whether or not the content and difficulty of questions accurately reflect curriculum standards. Scores are compared to previous examinations, and adjustments may be made to control for increases in the level of difficulty.

Thus, across all the steps just described, teacher involvement in testing appears to be much more extensive than is common in U.S. programs. For example, in 1990-91, 500 Alberta teachers were involved in developing, piloting, or serving on scoring panels for one English examination taken by 24,000 students. (In other words, one teacher was involved for every 48 students targeted for the examination.) In the United States, by contrast, New York—which leads all other states in the number of tests developed by the education department and the number of teachers routinely involved in test development—about 46 teachers are hired to develop each Regents Examination and about 10-20 teachers prepare other state-developed tests, which are taken by over a million students.[13]

The Alberta Assessment System

The procedures used to develop the assessments generally parallel those used to develop the examinations—beginning by reviewing the curriculum, making iterative reviews to reach consensus on the specifications, and field-testing of draft items—all with the maximum feasible paid teacher participation. However, unlike the high school examinations, the assessments measure achievement in core subject areas, such as science or social studies, at various grade levels (usually grades 3, 6, and 9) and each major subject area is assessed every 4 years.[14] In addition, cutoff points to describe levels of performance are determined differently. After teachers score the assessments, other teachers review them to set cutoff points—one for standard, or acceptable, performance

[12]Markers, like item writers, are selected by nomination. Marking takes place full-time, after each semester in which an examination is given. Markers are paid an honorarium of U.S.$51 a day plus expenses.

[13]Written sections of the New York State Regents Examinations and other tests are marked at each school by teachers as part of their regular duties in contrast to the practice in Alberta and British Columbia where teachers are brought together for training before and during the marking. However, the New York Education Department surveys all teachers of Regents courses after the examinations are administered to gather information on their reactions to the tests and suggestions for the future.

[14]Steps believed important to the quality of tests have been preserved even in times of tight money. For example, Alberta officials explained that they would assess knowledge in fewer grades rather than drop field-testing of the instruments or centralized scoring.

and another for excellent performance. These results are then compared to the actual student scores to determine if adjustments in scoring standards are needed.

The assessments also use a broader set of methods to measure student performance. The assessments used during the 1992-93 school year included activities that could be scored by teachers to measure oral communication, group participation, and problem-solving skills, in addition to multiple-choice and written-response items. These performance items are believed to measure skills that cannot be captured by paper-and-pencil tests, and they also serve to educate teachers by demonstrating how complex skills can be taught and measured.

The assessments are not intended to provide information for use in student placement or evaluation, but are designed to portray general patterns of groups of students' achievement to help improve teaching, direct revisions of the curriculum, and provide public accountability by showing attainment of the provincial standards.[15] Therefore, the advantages of the performance items in increasing the content validity of the assessments and in demonstrating new targets for teaching and new ways of testing are believed to offset their difficulties. (Problems include demonstrating the reliability of the items for an individual student, problems in verifying observational ratings, and high administrative and scoring costs.) Also, again because the tests are not used for decisions regarding individual students, sampling strategies are used to limit the costs associated with the performance items. The performance parts of the assessments are typically administered to a small representative sample of between 500 and 600 students in targeted grades. Machine-scorable items are taken by all students in targeted grades.

Alberta officials are considering changes to the assessments. They plan even broader involvement, such as bringing the community into test development by convening committees of parents and professionals working in selected fields to review and advise on the assessment specifications. Also, they are considering the feasibility of developing integrated assessments that comprehensively measure achievement in more than one subject at one time to shorten the time between assessments in a subject.

[15]On safeguards such as delayed reporting of results to forestall misuse of assessment results for evaluating students or teachers, see the discussion in chapter 3.

Testing in British Columbia

British Columbia is Canada's third most populous province, with about 3 million people.[16] More than 70 percent of the population lives in urban areas clustered about the southwest border of the province. As in Alberta, all grade 12 students are required to take at least one examination in either a college preparatory or general language arts course to graduate. British Columbia students enrolled in any of 15 courses are required to write examinations (compared to 9 in Alberta).[17] Students seldom enroll in more than two or three of these courses. The number of students taking the tests varies greatly; in recent years, about 28,000 took the academic English exam while only about 20 took the Latin test. In the second type of testing, the British Columbia Ministry of Education regularly assesses student achievement in English, mathematics, science, and social studies at three different grade levels; other subjects are assessed at the discretion of the province.[18] Core subject areas are tested every 4 years. Beginning in 1991, only a sample of students took the assessment tests.

The British Columbia Examination System

In British Columbia as in Alberta, officials explained that examinations are developed with broad teacher involvement, although some different groups and procedural steps are involved. There the teachers' federation plays a key role, items are included based on professional consensus and are not field-tested with a student population but by teacher teams, and oversight authority is assigned to a politically appointed independent provincial board of examiners that typically includes educators from the university and public school community.

Test development begins with an examination of the curriculum to determine which areas of knowledge and skills are most important. The minister of education and the teachers' federation select panels of teachers to develop the specifications for each examination. Approval from the board of examiners completes the specification stage.

[16]British Columbia's 500,000 public school students are roughly as many as attend public schools in Mississippi.

[17]Exams are prepared for biology, chemistry, communications (basic-level English), English (academic-level), English literature, French (for English-speaking students), French (for French-speaking students), geography, geology, German, history, Latin, mathematics, physics, and Spanish. French translations of exams are prepared at the request of French-speaking students. New exams are developed each year, with different forms available for each administration.

[18]Students in French immersion programs and French schools are assessed, at three grade levels, in French.

Teams of experienced teachers next write examination items, working part-time, after the regular school day or on weekends.[19] During the summer, team members meet with test development experts from the provincial education ministry to assemble different forms of each examination.[20] After the draft examinations are assembled, committees of teachers check them to be sure they match the provincial curricula and make recommendations to the original item development committee for study. Soon after the recommendations are received, the two committees meet to reconcile differences. To maintain security, the tests are not field-tested with students. However, after the examinations are revised, two teachers who have not seen the exams serve as trial writers and take each test as a further check to ensure item clarity. Difficulties that surface during the trial writing result in further revisions. Disagreements within or between committees are arbitrated by the board of examiners, which has final authority to approve the examinations.

As in Alberta, large numbers of teachers have participated in the examination process. For example, officials estimated that 80 percent of all teachers of examination courses in English in the province have had the experience of either developing items or grading examinations.

The British Columbia Assessments

In British Columbia, activities that guide the development of the assessments are more top-down than in the case of the examinations. The work depends less on teacher leadership and instead involves contract teams (usually directed by university-based subject-area specialists and composed of school teachers and graduate students) to design specifications, develop items, and score open-ended questions. As in the case of the examinations, the assessments reflect provincial curricula. For each subject assessment, an advisory committee of public school teachers, school administrators, and university professors guides the contract teams in creating the test specifications and in writing assessment items. Items are field-tested with a student population, and review panels of teachers

[19]Teachers are paid an honorarium of about U.S.$85 a day plus travel expenses for work performed during the weekend and the summer and U.S.$15.50 per hour plus expenses for work done after school.

[20]In British Columbia, as in Alberta, different examinations are used each semester. Three forms of the examinations are prepared to accommodate students in schools offering courses over one semester or during the summer session.

evaluate the appropriateness of the questions.[21] All stages of the activities are directed by the provincial education ministry.

Usually, all students in grades targeted for the assessments take part, though each writes only part of the test. A sample of students drawn to represent all parts of the province takes the performance components. A new assessment is developed for each test administration.

Plans call for the development of an expanded number of assessments, including assessments in fine arts, practical arts, and health education. In addition, the ministry plans future assessments to measure communication skills and problem-solving skills every 2 years, as well as other skills that are taught across several subjects.

Summary and Conclusions

Answering our first two evaluation questions, we found that in Alberta and British Columbia, educators have set curriculum-based standards for what should be taught at the provincial level and built on that foundation two very different types of tests for two measurement purposes. Considerable care is taken in selecting which material to test and performance standards to be applied when judging students individually. The procedures for both types of tests include extensive participation of teachers and multiple cycles of review and revision as tests are made as well as scored. Yet this care and thoroughness in validation and review do not preclude the development of new tests every year.

Both Alberta and British Columbia developed new curriculum-based tests in the 1980s. By contrast, we found in another study that only half the states reported they even had a curriculum framework in place.[22] Especially important is the wide teacher participation we found in these provinces that can be useful in setting a curriculum, developing sound tests based on it, and building local ownership of, and support for, the tests. (Public and professional opinions of provincial tests are discussed in chapter 4.)

[21]Officials of the British Columbia education ministry told us they regarded item security as less important on the assessments than on the examinations because assessment results are not used on the individual level. They conclude that the advantages of field testing with students to improve the tests outweigh the potential loss of complete security.

[22]Our survey of state testing practices, including their relation to a curriculum framework among many other issues, is reported in Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination (GAO/PEMD-93-8, January 13, 1993).

The experience of the provinces also shows the trade-offs they faced in quantity, quality, and costs of tests and the decision typically reached to test fewer students, in fewer subjects, and less often in order to gather more costly but better information about student achievement.

The efforts to define curriculum and measure results in these provinces appear to have been largely satisfactory in that there has been little pressure for national standards and tests as in the United States. The efforts are also consistent with the regional identity of "Canadians," citizens of a country with a small population spread over a vast area with divisions based on language and geography. (A small-scale, recent development to measure achievement Canada-wide is described in chapter 5.)

Finally, the official testing policy in Alberta and British Columbia distinguishes different purposes for testing and involves separate methods for examining individual achievement in critical courses and for checking how well students in general are achieving in broader areas of learning. Discussions in the United States often suggest a single test for both student appraisal as well as system monitoring, while the Canadian experience suggests the provinces have not found this desirable.

Chapter 3

Stakes and Safeguards Associated With Canadian Tests

This chapter answers the third and fourth evaluation questions addressed in this report: What types of stakes have been attached to tests to ensure that they will be taken seriously, and what explicit safeguards are used to prevent misuse of tests? These two questions arise from very different views of testing.

On the one hand, there are questions about the stakes that should ride on any test. Those who believe tests can be a valuable influence for stimulating teacher and student effort commonly propose that test scores be used to help make important decisions, such as high school graduation, college entry, or job placement. That is, they argue that stakes on at least some tests should be high, so that attention is paid. This argument is made by many groups advocating the establishment of national standards and tests, such as the National Council on Education Standards and Testing (NCEST). Canada has some experience with this approach in its examination system in five provinces.

On the other hand, those more skeptical about tests—that is, those concerned about their technical shortcomings, the difficulties of accurate measurement, or their users' misunderstandings—want more safeguards against flawed testing situations and against misuse of results. These safeguards are needed, they argue, to protect individuals by ensuring that testing is as fair as possible in the first place and that results are given only the limited uses they can sustain. Critics of proposals for U.S. national testing frequently cite these concerns. Canadian provinces have taken some careful steps to prevent this sort of misuse.

The United States has had much experience with the use of educational testing for high-stake purposes. Many states have implemented "gatekeeping" tests which determine student eligibility for promotion, while others use tests to determine eligibility for school entrance or placement in differentiated educational tracks. Some colleges require students to achieve scores above specified thresholds on standardized achievement tests before they can be considered for admission, and the military uses a standardized battery to determine both eligibility for the military and suitability for particular training programs.

Despite this tradition, the use of tests for high-stake purposes has remained controversial. First, the equality of opportunity to prepare for tests is a concern, since different student populations receive unbalanced shares of educational resources. Second, test developers have been slow to address the complexities of designing items that fairly test the

performance of members of different gender, ethnic, and racial groups without bias. And finally, many tests have been used for purposes that are not valid and have had consequences for both individuals and the educational system that many consider negative.

In this chapter, we address how the Canadian provinces have balanced the need to attach stakes to tests, so that tests and the curricula they represent will be taken seriously, and the need to protect students from unfair testing practices and unfair uses of test results.

Examination Stakes

At the time of our review, 5 of Canada's 10 provinces required high school students in the highest required course in a sequence or the final course in selected subjects to write a provincewide examination. (We discussed in the preceding chapter the procedures for developing and scoring these exams.) These requirements apply with considerable variation across the five provinces; for example, 18 courses have such exams in Newfoundland, but only one in Manitoba.

The specific stakes involved are both direct and indirect. Direct stakes include the requirement in all five provinces that students' final grades reflect the exam results to varying degrees (from 30 percent to 50 percent of the grade rests on the exam). Further, universities are encouraged to use the exam results in admissions and placement. (Canada has no uniform entrance examination such as the Scholastic Aptitude Test—recently renamed the Scholastic Assessment Tests—widely used in the United States.) Less directly, parents and schools can monitor whether teacher and student attention and effort are directed to relevant material. And finally, just as in the United States, comparative rankings of examination results among classrooms, schools, or districts can be used for formal accountability (though this is generally regarded as inappropriate) or for overall policy guidance.

Stakes for Students

We found evidence, in the student stakes assigned, of conflicting evaluations of the examination systems in these provinces and at least some second thoughts about their relative importance. The highest stakes (requiring that 50 percent of the student's final course grade must rest on the exam and that the other 50 percent, awarded by the teacher, be adjusted on the basis of exam scores) are found in Quebec and Newfoundland, the two provinces that have the longest running exam systems and never abandoned them when the rest did in the 1970s. In

these provinces, concern regarding the comparability of marks assigned by different districts, schools, and teachers is so high that teacher-assigned grades are adjusted statistically based on how each class does on the exam. (For example, a class that scored unusually low on the examination but which received unusually high grades from its teacher could have those high grades reduced.)

In contrast, in provinces that have more recently reestablished exams, either the stakes are not quite so high or else they have actually dropped over time to reflect the greater importance those provinces give to teacher judgments of classroom work. Also in these provinces, teacher grades are not adjusted (as just described) following review of the exam results, because differences are considered normal. That is, teachers are thought to be aware of student performance in more areas than can be covered or evaluated by one paper-and-pencil test (though Canadian observers warned us there is no evidence concerning how teachers assign their share of students' grades). Thus, in Alberta, teacher grades count 50 percent and are independent; further, proposals are being considered to decrease the weight of the standard exam to 40 percent. (The remaining 60 percent would be made up of a combination of 40 percent teacher grades and 20 percent teacher evaluation of student work on province-set performance tasks.) British Columbia has reduced the exam share of a student's grade from 50 percent to 40 percent. The most recently established exam program, in Manitoba, required no more than 30 percent of a student's grade to rest on the new test.

Linking test results more tightly to students' fate in later college or job decisions is often mentioned in U.S. policy debates as a possible stake that could make students take school work more seriously, and provincial education officials told us that they had just such intentions. Yet examination results, either specifically or when blended with teacher marks to form final grades, are not especially visible in either postsecondary school or job decisions. In the university admissions case, timing makes it difficult: the June results are too late to make any difference for most students, although they may be critical for those with borderline eligibility status. As for the use of results in hiring, the business leaders and educators we spoke to told us that examination scores are not commonly scrutinized. Survey data in two provinces confirmed that, as in the United States, employers check with schools only on the most basic questions about an applicant's graduation and skills. The exam results (along with many other data) do get used, we found, in decisions within

the Canadian military concerning candidates for officer training, but this affects only a very small number of young people.

Other Uses of Examination Results

The examinations' importance may actually be greater than that suggested by the mandated uses. We heard repeatedly of the high visibility of the examinations as a key indicator of the overall state of student learning in a classroom, school, or district. In some provinces, results are used for program evaluation. The simple exam scores and rankings of schools or districts on average scores draw wide press and public attention, much as scores on college admissions tests do in the United States. This continues to be true even though most Canadian examinations are taken only by students who have passed prerequisite courses and are thus not representative of the general student population in a class, school, or district.

Other Canadian uses of examination results for formal accountability or for making classroom assignments can make the stakes high for teachers. Teacher groups in the United States have consistently opposed such uses of tests; data on the acceptability of testing when used in this way in Canada also show teacher opposition.[1] In addition, provincial documents routinely warn the public against using exam results to judge educators or schools.[2] Some go on to describe other influences on test results beyond the quality of teacher and student effort—basic demographic differences, for example, of language or economic status—that are also related to school success. Still other influences on test results are differences across school programs such as student selection procedures (initial screening of class groups through formal or informal prerequisites for admission, or exclusion of learning disabled students from exam classes) and other factors.

Testing critics in Canada, as in the United States, believe that inappropriately high stakes given to exam scores can result in educational inequities. For example, we heard that "marginal" students in Canada are commonly discouraged from taking exam courses (or encouraged to drop

[1]A survey by Calder (1990) found that two-thirds of teachers surveyed in Alberta opposed the use of examination scores to evaluate teachers. A survey by Anderson et al. (1990) of teachers in British Columbia reported that, on average, they viewed the use of examination results in the evaluation of teachers as detrimental.

[2]For example, such cautions are routinely included in the following annual publications: Annual Report: Diploma Examination Program, Alberta Education; Report to Schools, Ministry of Education, British Columbia; Educational Indicators, Department of Education, Newfoundland and Labrador; and Provincial and School Board Results, Ministry of Education, Quebec.

them if they do poorly) so as not to depress a class or school average. We were also told that it is common to assign the strongest teachers to classes aimed at preparing students for the exams. We found little discussion of these practices in the Canadian testing literature, however.

Safeguards for Students in Design and Scoring of Examinations

Most of the five provinces' examination systems have developed various sets of safeguards against misuse of test results. These include:

- linking the tests to what is supposed to be taught and taking steps toward providing all students with more equitable opportunities to learn material needed for success;
- adapting the testing situation for students with special needs;
- promoting reliable scoring;
- providing multiple opportunities to retake examinations; and
- minimizing unintended increases in the difficulty of examinations from one year to another.

Many of the provisions that work toward increasing the fairness of the examinations are built into the test development process discussed in chapter 2. These procedures include the development and dissemination of curriculum-based examination specifications, widespread iterative reviews of examination items, and broad-based teacher involvement in all aspects of the examination process. Targeted mainly at teachers, the procedures are intended to increase the content validity of the examinations and increase the alignment between what is tested and what is taught. In addition, several provinces routinely provide students with copies of examination specifications, study guides, and sample examinations to help them prepare.

Especially important in promoting equal opportunity for students in Canada is a factor outside the examination system itself. Provincial funding formulas tend to level resources among schools within a province and are a step toward ensuring that each school has comparable resources to implement the curriculum. Thus, the funding formulas contribute greatly to the fairness of the conditions under which students prepare for the tests.[3]

Other safeguards address the individual needs of students who are disabled, in crisis, members of a linguistic minority, or otherwise at risk of

[3]On average, provincial government grants account for about 70 percent of net general expenditures on public education. (Government of Canada, Education in Canada, External Affairs and International Trade, Canada, 1989.)

not testing as well as they might. Examinations may be modified to accommodate students with physical or learning disabilities at the discretion of a board of examiners or provincial education agency. Procedures exist to excuse students from writing examinations because of illness, accident, or bereavement. Examinations are prepared in both English and French.

We noted another safeguard, aimed at avoiding inappropriately high levels of examination difficulty. Where a specific subject is a general graduation requirement for all students, the provinces with examination systems, with one exception, offer different levels of examinations.[4] In British Columbia and Alberta, for example, different examinations are tailored for the basic and academic English courses, both of which satisfy a requirement for a senior year language arts course. Also, as we discussed previously, participation in most examination courses is voluntary except for college-bound students who have to meet college admissions requirements. Thus, students who are at risk can avoid meeting requirements that may be insurmountable, and students who are talented academically avoid instruction that may lack challenge. These practices, however, tend to reinforce stratification between high school students in academic and nonacademic streams. In addition, these practices also reduce the comparability of examination results; the proportion of students enrolled in different examination courses changes from one year to another, which diminishes the potential usefulness of aggregate results in monitoring academic achievement over time (though that is not a primary purpose of the examination system).

Canadian provinces take a number of precautions designed to safeguard against idiosyncratic grading practices. In Alberta, British Columbia, and Manitoba, all open-ended questions are graded independently by two or more teachers. Statistical checks identify disagreement among grades assigned each item by different raters and trigger procedures for further independent ratings to reconcile differences. In Quebec, the provincial education ministry rescores a sample of examinations in a number of subjects from a sample of schools to monitor the accuracy of school-based teacher scoring. If significant differences appear between school and ministry grades, all the examinations from the school are rescored by provincial staff.

[4]This is not applicable to practices in Quebec where there is no streaming of students in academic courses.

Four of the five examination systems have formal guarantees that students may obtain a second scoring of examinations. For example, in Alberta, students pay a fee of about U.S.$20 for rescoring. The fee is refunded in these provinces if the rescoring results in an upward score adjustment. The fifth province (Manitoba) has no explicit procedures for rescoring but has an unofficial policy to respond to concerns of individual students by rescoring examinations.

All provinces with examination systems allow students to retake exams. Students may rewrite examinations two or three times without repeating a course, or students may elect to repeat a course and then rewrite the exams. The highest grade achieved on the examination is recorded on a student's transcript.

A final set of procedures is designed to prevent students from failing the examinations because their difficulty may have increased from one year to another. New exams are created each year to safeguard the content of the tests, and as test standards are redefined and new test items are written, their difficulty levels may also change. To compensate for substantial changes in difficulty, provinces adjust scores to stabilize pass rates. In British Columbia and Quebec, grades on examinations are derived from scaled scores rather than raw scores, with scaling based on judgments regarding the proportion of grades that should fall at different intervals. The Newfoundland education department has traditionally employed scaling techniques that allow upward shifts in scores.[5] In Alberta, cut-scores are adjusted to prevent substantial fluctuations from one year to another.

Assessment Stakes and Safeguards

In contrast to the examinations, provincial assessments are not intended to have stakes for individual students, although the results may have consequences for schools and districts. Also, the provincial assessments are more diverse in terms of format and student and subject coverage than the examination programs and, therefore, so are the policies and procedures to safeguard the use of these tests. In general, however, all eight provinces with assessment programs established them to obtain estimates of the performance of students as a group and indications of

[5]New processes used by Newfoundland in developing a biology exam attempt to improve the comparability of examinations in measuring achievement in a particular course from year to year. First, information from field tests can be used to predict changes in achievement patterns as measured by certain test items over time. Second, item banking makes it possible to gather longitudinal information on response rates to particular items, thus making it possible to detect changes in the performance of different student cohorts.

program success rather than to obtain information for use in educational decisions about students and teachers. Safeguards to further this intended purpose of the test include:

- the development of procedures to make sure the full student body takes part in the assessment testing or is eligible for sampling (i.e., that participation is not manipulated);
- the use of sampling among students to maximize test coverage, reduce costs, and reduce inappropriate use of scores at the individual level; and
- the aggregate and delayed reporting of results to eliminate misleading comparisons.

Several provinces have procedures that make it difficult for schools to artificially inflate scores by somehow keeping weaker students out of the assessments. (See the critique of such manipulation noted earlier in this chapter.) For example, in Alberta, each superintendent of schools must approve all students excluded and must send the provincial authorities a list of all those omitted and the reasons for omission. In British Columbia, all students enrolled in a program of studies are targeted for assessment, with exclusions made only on a case-by-case basis. Other safeguards, such as delayed and aggregated reporting of scores and the use of sampling techniques, have decreased the exclusion of marginal students by reducing the stakes schools and districts may attach to assessment results.

Five of the eight provinces with assessment programs test only small samples of students in order to afford the cost of assessments that include open-ended or performance items requiring individualized administration and scoring. Although these open-ended items are more costly than large-group tests with machine-scored items, most of the Canadian test experts we spoke with agree the inclusion of these items is necessary to measure certain important standards. In Alberta, although all students in targeted grades complete the multiple-choice and written portions of the tests, only a sample of students complete the performance items that may require oral communication, manipulation of equipment, or group participation skills.

Beginning in 1991, the British Columbia education agency moved from assessing all students in targeted grades to assessing stratified samples of students, with subsamples of students completing performance items. In Manitoba, samples of about 10 percent of students in grades targeted for the assessments are tested. In Ontario, all high school students enrolled in targeted subjects and samples of 100 English and 100 French elementary

schools complete the machine-scored parts of the assessment; performance items are administered to subsamples. In Quebec, the assessments involve student samples selected as needed.

Although sampling was implemented largely as a cost-saving measure, the implementation of sampling strategies is associated with improved testing practices. First, sampling has made it possible to include complex open-ended items that not only improve the information measured by the tests but also have the potential to increase the value of the tests as instructional exemplars and thus encourage teaching of higher level skills and demonstrate to teachers how these skills might be measured. Second, sampling strategies impede the use of assessment results in decisions regarding individual students and teachers, thus safeguarding individuals from unintended use of test results.

Four of the eight provinces with assessments report scores only on the provincial level to prevent comparisons among teachers, schools, or districts—comparisons that are generally regarded in Canada as simplistic and are very unpopular among teacher organizations. In addition, Canadian officials hope that aggregated reporting will cut down the incentive to inflate scores by excluding weaker students and overteaching what is tested. The provincial education authorities in Manitoba, New Brunswick, and Quebec have routinely reported results on the provincial level only. In British Columbia, results had been reported on the school, district, and provincial levels, but beginning with the 1991 science assessment, scores were reported on the provincial level only.[6]

Although Alberta reports scores on individual, school, and district levels, the reporting of scores is delayed to deter their use in student placement decisions and in teacher evaluations.[7] The assessments are given during the spring semester, and scores are not reported to schools until the following October.

[6]The new reporting procedures coincided with changes in the format and content of this assessment, which was developed from specifications that differed substantially from previous ones and included new types of performance items. As a likely result of these factors, scores on the assessment were lower than expected, which increased teacher opposition to disaggregated reporting. In addition, sampling strategies were used to offset the costs of the performance items, which in turn reduced the reliability of scores at the school and district level and thus the desirability of reporting at these levels.

[7]Individual scores reflect performance only on multiple-choice and written-response questions.

Summary and Conclusions

Answering our third question, we found that officials in the Canadian provinces attach stakes to some, but not all, provincial tests. Examinations carry substantial stakes for high school students, teachers, schools, and districts. In the five provinces where examinations are administered, students' examination grades determine between 30 percent and 50 percent of their final course grade. Comparative rankings of classrooms, schools, and districts are frequently used by educational officials and the general public in making judgments about the relative quality of instruction that is being provided in different settings. In contrast, the provincial assessments are not intended to have consequences for individual students. Canadian officials have not found it necessary to attach stakes to these tests to emphasize their importance and have developed procedures to prevent the use of the results in decisions regarding students.

The stakes attached to the high school examinations have the effect of increasing the comparability of student grades, underscore the importance of the curriculum framework and standards on which the exams rest, improve efforts devoted to teaching and learning, and encourage accountability for results. Yet in the interest of fairness, the stakes are limited through the use of blended scores.[8] Other practices, such as setting exams only in certain subjects and providing alternative graduation tracks, exclude large groups of students from these potential benefits and decrease the value of the examinations in monitoring provincial achievement (though the latter is not their main purpose). However, this trade-off allows the examinations to reflect standards that are academically challenging and measure high-level achievement without creating an unacceptably large pool of failures and drop-outs.

Concerning safeguards, our fourth question, we found many. Examination safeguards have been developed to allow students multiple opportunities to retake examinations, to request that an examination be regraded, and to accommodate students who are disabled. Safeguards associated with the assessments, such as testing only samples of students and reporting results in aggregated forms and after a delay, address the need to provide accurate and comprehensive measures of provincial achievement while protecting students and teachers from invalid uses of assessment scores. However, these same safeguards prevent the use of the assessment tests to identify students who may need additional help or to give timely feedback

[8]As already noted, it is an open question whether provincial tests or teachers more "fairly" assess student learning. We are reporting the policy views of Canadians who believe fairness requires a blending of both sources of information; some Canadians also believe one or the other should be relied on more heavily.

to individual teachers (though again, these are not the assessments' purposes).

Two safeguards in particular highlight the differences between the Canadian and U.S. situation. These concern equitable delivery of the curriculum and involvement of teachers at all steps of testing. First, the role of the provinces in funding education has enabled school districts to provide more nearly equitable opportunities for Canadian students. Unlike the situation within some states, testing may be seen as fairer in the absence of gross differences in educational resources from one Canadian district to another in a province.

Second, widespread teacher participation in test development and in centralized grading have increased the reliability of test scores and the likelihood that teachers know, in general, what will be required of students. These same steps have increased examination costs in dollars and in time. The increases in cost, in turn, have prevented provinces from expanding the number of courses with examination requirements. However, widespread teacher participation is increasing because of its apparent advantages in improving the quality of the tests. These advantages are recognized more fully in Canada than in the United States, where teacher involvement and the use of grading panels is both more limited, in terms of the numbers of teachers, and less prevalent, in terms of the number of states that routinely use teachers in test development activities.

Chapter 4

The Promise of Testing for Improving Teaching and Learning

Support for a U.S. national achievement test or testing system rests ultimately on the belief that improved tests, with appropriate stakes, will lead to improved achievement. Our last question, concerning the perceived effects of tests on teaching and learning, reflects the hope of advocates of expanded U.S. testing that such effects will occur, but there is no precedent for such an effort in this country, hence no evaluation evidence that can be weighed in the discussion. Knowledge about the Canadian experience could thus be useful in U.S. testing debates.

In this chapter, we review what we have learned from interviews, documents, polls, and other studies about how testing is believed to have affected what is taught and what is learned. We found no direct empirical evidence of the effects of tests on teaching and learning. We limited our review to evidence from Alberta and British Columbia because their efforts have been more noticeable in the last decade as they reestablished previously abandoned examinations and because their testing efforts include approaches like those advocated for the United States.

Observed Changes in Teaching

We found general agreement that both the examinations and the assessments had affected what is taught and how. Testing inevitably involves selecting only certain content and skills to be examined from the broad array of schoolwork in a grade or subject in a year. Teachers reported that the provincial examinations have the effect of getting them to spend more time on topics included in the published specifications and less time on discretionary activities.[1] Concerning tests' effects on teaching methods, we saw reports that high school teachers in examination courses emphasize transmitting information (through lectures or worksheets) they believe will appear on the tests, at the expense of group participation, discussion, or problem-solving activities that are more common in nonexamination classrooms.[2] Evaluation of these outcomes is controversial; that is, these effects of testing on content and methods can be viewed as a welcome assertion of priorities and standards that bring

[1]Anderson et al. (1990), in surveying 947 British Columbia teachers, found that, on average, they report spending more time coaching students for tests and on test-related activities and spending less time on authorized supplemental materials as a result of the provincial tests.

[2]For example, in their study of the impact of grade 12 science assessments and examinations on the instructional activities of teachers in British Columbia, Wideen et al. (1991) found lecturing increased while laboratory work and discussion decreased at upper grade levels when grade 12 students were targeted for provincial tests. Thus, grade 12 science teachers spent 57 percent of their time lecturing or doing other teacher-centered activity compared to 32 percent for the grade 10 teachers and 15 percent for grade 8. Alternatively, grade 8 teachers spent 24 percent of their time on laboratory activities; grade 10, 8 percent; and grade 12, 7 percent. About 10 percent of instructional time at grade 8 was spent in discussion compared to 4 percent at grade 10 and 12.

coherence to classrooms or as an unfortunate narrowing of teachers' autonomy and of variety in teaching.

Provincial education officials in both Alberta and British Columbia told us that they believe the tests have indirectly led to improved teaching through the many training opportunities that center on them. For example, teachers developing the tests and grading them are involved with ministry officials and university experts in sustained conversations that can lead to consensus about what to teach and what learning to expect from students. Teachers are provided with opportunities to learn new ideas about how to measure student learning and to practice the skills involved in making reliable observations of student performance through their involvement in developing the innovative assessments being tried in the Canadian provinces. We heard everywhere we visited that teachers eagerly pursue the opportunities for involvement in provincial testing activities.[3] Provincial officials believe that those teachers who are selected are effective in increasing the credibility of the tests and in communicating current ideas about what to teach and how best to measure student learning to teachers in their school districts.

Observed Changes in Student Learning

Neither Alberta nor British Columbia has any independent measure that can show trends in student learning over time to help answer questions about whether establishing testing programs had any noticeable effect. (We discuss in chapter 5 a new plan to develop a Canadian national test that could provide such a measure in the future.)

Student surveys showed that students in examination classes noted, as did their teachers, the narrowing of what is taught and the lack of time for material unrelated to the tests.[4] Students reported that the exams increased their motivation, which in turn may cause more effort and learning, but they also reported pressure and anxiety, which may hinder learning.

Indirect evidence suggests examinations have influenced the amount of time students spend in high school. According to provincial officials, reintroduction of examinations in both Alberta and British Columbia coincided with increases in the fraction of students staying in high school

[3]Prestige and professional collegueship were commonly cited in interviews as motivating teachers' participation; but extra pay was probably also important. For example, teachers from one part of Alberta stopped participating in item development in 1990-91 when the province did not offer pay; they resumed participation when pay was restored.

[4]Survey data were reported in Anderson et al. (1990) and Calder (1990).

for an additional year; about 30 percent of students in Alberta and 10 percent of those in British Columbia now do so, and we were advised that the number is as high as 40 percent in some schools. Interpretation of such figures is again controversial, and we noted competing arguments over whether students' time and effort in such an extra year resulted in important learning gains or unfortunate focus on narrow examination material in order to raise a test score. In any case, the students who delay graduation represent a notable dollar cost to the education system as well as opportunity costs to individuals and society.

The exams are also considered to contribute to increased separation of groups within schools. We heard claims that the establishment of the examinations (and the visibility of the results) confers on those classes a higher status for the students and teachers involved and that students can be increasingly isolated from each other to the degree that lower ability students are discouraged from enrolling or, if enrolled and doing poorly, are encouraged to drop the class so that the class average score on the exam will be as high as possible. To this, however, some objected that grouping occurs anyway (as students group themselves socially and since schools in general prefer to group students of like aptitude for schoolwork) and thus, establishing any unique stratifying effect of the examinations is hard to do.

The argument over this perceived effect of the examinations involves the long-standing general debate between those who believe that a school's academic mission can be enhanced by "tracking" or streaming subgroups of students and others who stress the social integration as well as the educational motivation that can come from common schooling of all students with fewer distinctions. We found no evidence that could help establish the key factual point of the degree to which examination courses are genuinely open to all; a generally established research finding is that high expectations for all students are important to drawing out their best.

Support for the Examinations

Teachers and the public in Alberta and British Columbia generally approve of the provincial examination systems and believe education has benefited. For example, a majority of the teachers surveyed in both Alberta and British Columbia favored retaining the exams, finding them well done and fair for students while judging their overall effect on the

scope of instruction as negative.[5] Perhaps most important, teachers and administrators believe the exams won public support for the Canadian schools. The general public, according to public opinion polls, is strongly supportive of the examinations and links them with higher educational standards.[6]

We were told, and research documents indicate, that university officials and members of faculties of arts and sciences generally favored the examinations, associating them with establishing higher educational standards and increasing comparability among student grades.[7] However, the members of education faculties that we interviewed were more critical of the examinations. While viewing them as successful in clarifying instruction goals and promoting a more uniform curriculum, these respondents in general faulted the examination system for overlooking the needs of nonacademic students, creating more stressful learning conditions for teachers and students, fostering a narrower curriculum, and reducing the importance of classroom work in the determination of final course grades (to the extent that this is the key factor in teacher grades).

Business officials we spoke with also strongly supported provincial testing in general. Despite the fact that employers did not appear to use the examination results, public support for more testing (both examinations and assessments) is plausible; that is, the public may be responding favorably to a symbolic signal of greater concern about the outcomes of education by provincial authorities and school officials or may approve the tests in hopes that educators will use the results to diagnose problems and make improvements.

Summary and Conclusions

We lack empirical evidence to answer our final question regarding the promise of tests to improve teaching and learning. However, we found a

[5]For example, Calder's 1990 survey of 135 teachers in Alberta found, on the one hand, that 65 percent of teachers polled indicated that the examinations should continue to be given, nearly 75 percent that they were well constructed, and 57 percent that they were fair to students. On the other hand, 60 percent of teachers surveyed believed the examinations were required more for political purposes than educational ones. In addition, Anderson et al. (1990) found over half the teachers he surveyed in British Columbia indicated the examinations had narrowed the breadth of courses and the effect of this narrowing was detrimental to learning.

We were told by researchers that teacher judgments of the examinations varied considerably, with math and science teachers more favorable toward the examinations than language arts and social studies teachers.

[6]Results of a 1985 provincewide survey conducted by the British Columbia Ministry of Education indicated 75 percent of the general public supported the reintroduction of the provincial examinations.

[7]Survey data were reported in Anderson et al. (1990).

consensus of opinion among the individuals we interviewed regarding the potent influence of testing on both what is taught and how it is taught.

Teachers, we were told, emphasize content that is likely to be on the provincial tests and provide instruction that is thought to fit the performance requirements called for by tests. On the positive side, teachers appear more likely to emphasize the priorities and standards of the provincial curricula. On the negative side, teachers' autonomy and ability to individualize the curricula may be constrained and instruction may be narrowed to conform to test requirements. A majority of teachers give acceptable marks to the tests' quality and fairness, but believe they are used largely for political reasons rather than for educational ones.

Students report the examinations have increased their motivation to study, which may increase learning, but also indicate pressure and anxiety about the examinations, which may hinder learning. Like teachers, students report instruction in examination classes is tied closely to the exams.

University officials and faculty members in the arts and sciences generally favored the examinations, linking them with higher educational standards and increased comparability of student grades. The members of education faculties we interviewed, however, criticized the examinations for overlooking the needs of nonacademic students, creating stressful learning conditions, narrowing the curricula, and decreasing the importance of classroom work in determining the final grade.

The general public and the business community support expanded testing and associate the introduction of provincial testing programs with higher educational standards.

Chapter 5

Steps Toward the Development of a Canada-Wide Assessment of Literacy and Numeracy

This chapter describes Canada's recent steps toward developing a Canada-wide national assessment. Efforts have been made to harmonize different provincial standards and curricula and build upon similarities to arrive at consensus on the specifications for a single national test. These efforts may show, on a smaller scale, what lies ahead if the United States forms regional clusters to develop new tests linked to standards as suggested, for example, by the National Council on Education Standards and Testing (NCEST).

Origins and Purposes of the National Indicators Project

The provincial ministers responsible for education in Canada, with the concurrence of their respective governments, established a Council of Ministers of Education, Canada (CMEC), in 1967. In the absence of any federal education department at the national level, the organization enables the ministers to act in collaboration on projects of mutual interest, to provide a liaison with various federal departments, and to represent Canadian education with other national and international educational organizations.

In September 1989, CMEC decided to investigate alternative approaches and costs for a Canada-wide assessment as part of a national indicator project. Such an assessment was seen as a response to growing interest in overall assessment of the effectiveness of provincial educational systems and also as a possible additional source of data to compare with that from the Third International Mathematics and Science Studies (TIMSS).[1]

The CMEC officially launched its overall School Achievement Indicators Program in June 1990, calling for comparative indicators of the performance of the educational systems and an assessment instrument to measure skills in reading, writing, and mathematics content and problem-solving of a sample of 13- and 16-year-old students. The CMEC stated the objectives of the indicator program as follows:

- to provide data that will assist each province and territory in making policy decisions, setting education priorities and planning program improvement, while respecting the autonomy of the provinces and territories in matters of education;

[1]The Canadian provinces have participated in many major international assessments, but TIMSS is the first that will include a representative sample of Canadian students at the national level. The student testing component of the TIMSS is scheduled to begin during the 1993-94 school year. In addition, students in Alberta, British Columbia, and Ontario will also participate at the provincial level.

- to collect information on the achievement levels of 13-year-old and 16-year-old students and on the participation and graduation rates to help determine the effectiveness of provincial educational systems; and
- to report in a manner that clearly informs Canadians of the information gathered concerning reading, writing, and mathematics and of the procedures by which the program was conducted, and also provide a possible information base for discussion leading to curriculum improvement at the provincial level.

The project focused on two kinds of indicators. First, statistics on student participation and graduation rates were collected and tabulated for each province.[2] This portion of the project is on-going, and data have been produced and published. The other part was to involve a new assessment covering reading, writing, and mathematics; this is the part that has proven more difficult, although not impossible, to achieve and that suggests implications for the United States.

Efforts to Structure the Project

After agreeing that a Canada-wide test would be advisable and financially feasible, the CMEC considered several ways of obtaining such a test. First, a technical committee was formed to review commercial and provincial tests and judge if they would suit the project. This committee concluded that no existing test was suitable. Another group was then formed to look into the policy and financial implications of developing new tests. This group recommended that CMEC advertise for an outside technical consultant to conduct the project, but after this plan was adopted and a formal request issued, response from the academic and consultant community was minimal.[3] Few proposals were received and all were rejected.

Still a third committee was formed to explore alternative methods to develop the new assessment and eventually proposed that a consortium do so, composed of the Corporation of the CMEC (a nonprofit corporation that provides the services of a secretariat to the Council) and the provinces of Alberta and Quebec, working along the lines of the plans

[2]Original plans called for the collection of additional information including retention rates, drop-out rates, and participation in special programs, but these elements were not retained.

[3]Some of the officials we interviewed suggested a number of explanations for this small response, such as: the commitment to other work of key members of the psychometric community; hostile views of the new test in the academic and consultant communities; and expectations that the work would be difficult to do because of the potential conflict among the provinces involved. (This indeed has proved to be the case.)

advertised earlier. The Council approved the formation of the consortium and appointed a policy advisory committee to oversee the activity.

As now established, the project thus involves the corporation, which provides liaison and communications with the ministries and the council, oversees the finances and funding, and calls and holds meetings relating to the project. The Alberta Department of Education has the lead role in carrying out the project; test development and subject-area specialists employed by that province have the responsibility for designing products for English-speaking students. Test development and subject-area specialists employed by the Quebec Ministry of Education, in consultation with Alberta, are responsible for designing products for French-speaking students. Provincial officials in Alberta and Quebec contracted with independent experts to provide advice as needed. The Ontario education ministry joined the program as a third member of the development consortium in December 1991 to coordinate the administration and marking of the assessment instruments.

Getting Started in Test Development—Standards

Initial issues for any test include what to test and how hard it should be. The purpose of the new Canada-wide test is to indicate reading, writing, and mathematical achievement. The consortium development team, in consultation with contacts in each provincial agency, drafted criteria describing five levels of performance in reading, writing, and mathematical content and problem-solving. Each province established its own procedure to react to these criteria. Following the provincial reviews, the criteria were revised to reflect as many as possible of the recommendations made by the provinces.

The next set of activities involved a series of interprovincial meetings and consultations with independent technical contractors. These meetings were held to develop consensus regarding the criteria, ensure common understanding of them, translate them into measurable tasks, and obtain provincial endorsement. The approved criteria were used to develop performance scales that measure a continuum of reading, writing, and mathematical competencies at five levels of achievement.

After the initial scales were developed, they were referred to provincial education agencies for consultation with various stakeholders including teacher organizations, trustee (school board) organizations, and parent and business groups in each province for review and further feedback. Thus, officials in each province had the opportunity to examine and, if

necessary, strengthen the match between the scales and the provincial curricula. After these activities were completed, a revised set of scales was generated to measure and describe 13- and 16-year-old students in terms of basic reading, writing, and math competencies.

Development of the Assessment Instrument

The next step was to produce the actual assessments to measure the levels of student skills. From the beginning, Canadian educators have accepted the need to develop new instruments to match the standards once they are procduced. (In the United States, by contrast, the National Assessment Governing Board (NAGB) has sought to set standards for basic, proficient, and advanced achievement based on the National Assessment of Educational Progress (NAEP), an existing test that was developed for other purposes. NAGB's approach has raised unforseen technical issues.[4] Canadians have shown greater awareness of the technical difficulties of their enterprise.).

Canadian teachers in a sample of schools in each province were asked to use the reading, writing, and mathematical competency scales to classify the performance of their students. Then, a common assessment, developed to measure students' performance in a manner consistent with teacher assessment of students, was administered to a sample of students. During the development phase, results on the assessment were compared to teachers' ratings of the same students to assess the validity of the test. The results of the analyses were used to modify items that did not produce results corresponding to teacher classifications. Taken together, this trial administration provided validation of the items used, the achievement levels, and the cutoff points between levels. A second trial administration was conducted to validate revisions in the instruments and the marking guides. This again is a departure from procedures used in developing the U.S. achievement-level standards intended to be measured by NAEP. In that case, NAGB did not adequately examine measurement validity issues.

The math assessment materials were approved by the nine participating provinces in December 1992 with administration scheduled for April 1993. The reading and writing assessment materials were distributed to each province for final approval early in 1993, with administration scheduled for April 1994.

[4]We discuss t his in an interim report, National Assessment Technical Quality (GAO/PEMD-92-22R, March 11, 1992). The full report of our work is forthcoming.

Reporting Results

Current plans call for reporting assessment results on a national and provincial level, with results disaggregated by gender, official language (both English and French in Manitoba, New Brunswick, Ontario, and Quebec), and age. Canadian officials believe the possibility of school and district reporting would compromise participation. In addition, any further analysis would involve more extensive sampling and increase the cost of the assessment.

Feasibility Issues

This ambitious program is in its third year of development. The program had an original cost estimate of about U.S.$1.3 million but is now expected to cost approximately U.S.$2.9 million over 4 years. Provinces agreed to contribute staff time of one person a year, plus printing and distribution costs, with contributions from the CMEC funding the remainder of the effort.

In reaching its present state of development, the indicators project encountered many conflicts, which observers believe will continue even though agreement has been reached to allow tests to be developed. A number of teacher and subject-area associations, particularly teachers of language arts, have demonstrated strong opposition to the project. These organizations contend that the principles on which the program is based are incompatible with sound teaching, classroom assessment practices, and the curricula of the different provinces.[5] They question the potential reliability of test instruments and doubt that results will accurately reflect what students can do.

Because of this teacher opposition, changes in the specifications for the assessments have been adopted along lines of the "more authentic testing" ideas widely discussed in the United States, including dropping plans to score the test entirely by machine and dropping a timed writing test. Instead, at least 50 percent of the reading assessment instrument will be open-ended, and examples of students' best or favorite pieces of writing will be collected and assessed to show how well students perform under regular classroom learning conditions. Adding these open-response and portfolio components may increase the cost by about U.S.$7-U.S.$8 per marked student paper, doubling the original budget for assessment administration and scoring.

[5]Some organizations contend standardized testing conditions cannot yield reliable results and therefore are against all forms of standardized testing. Other groups oppose methods of assessment that do not include extensive student collaboration and involve several days of classroom work.

Other issues were in dispute, largely resulting from the accumulated experience of separate provinces in developing curriculum standards and the tests linked to them. That experience has yielded integrated systems of standards, curriculum, and assessment with substantial teacher support. The Canada-wide effort is at risk to the degree that provinces must give up hard-won consensus. Thus differences on content coverage have already resulted in some disarray: officials in one province (Ontario) adopted observer status and returned as a participant only after a detailed memorandum of understanding was agreed to and adopted, and one province (Saskatchewan) has definitively decided not to participate in the administration of the mathematics assessment.

Consensus-building has not been as contentious among teachers of mathematics as it has been among teachers of language arts. More agreement regarding standards exists among mathematics teachers and this is reflected by a fairly high degree of symmetry among the mathematics curricula in most provinces. Most provinces have accepted the standards developed in the United States by the National Council of Teachers of Mathematics (NCTM), which represents American and Canadian mathematics teachers and specialists.

Broader consensus may also be a problem. Members of the business community appear to be very supportive of this effort, but some educators are less favorable. This is certainly related to the different views of education held by the two communities. Members of the Canadian business community believe the Canadian assessment should emphasize items measuring functional work-related skills such as computer spreadsheet literacy, while most teachers stress the need to emphasize high-level thinking skills such as creativity, and some do not believe wide-scale testing is helpful at all.

Summary and Conclusions

Despite the experience of the Canadian provinces with criterion-referenced testing, early steps in a project to develop a Canada-wide assessment of language arts and mathematics have been difficult. Similarities exist across provincial curricula; however, standards are defined and presented differently. In addition, the provinces have different preferences regarding assessment methods and the relative importance of different skills and knowledge. CMEC has been responsive to teacher concerns and has taken actions to accommodate teachers even though these actions have added substantially to the cost of the assessment.

These Canadian difficulties have occurred even though all agree the assessment is a low-stakes instrument. If it was to be used for high-stake purposes, such as decisions about individuals, consensus could be even more elusive. Other indicators, such as participation and graduation rates, will provide a context in which assessment results can be understood and monitored over time. Although Canada has but 10 provinces and one-tenth the population of the United States, the officials we spoke with agreed that the differences among provincial curricula were too great and important for there to be high stakes or consequences riding on the results of the new test. If our states should develop an equivalent commitment to curriculum standards and produce regional tests to measure student attainment of those standards, we believe that a broad national test—even one with low stakes—emerging from the harmonization of those regional efforts could encounter similar conflict.

Conclusions

Our Research Questions About National Tests

In this report, we have examined Canadian school achievement testing because that experience may shed light on issues currently being debated in the United States about national tests as part of a broad school improvement effort. The five questions organizing our study were:

1. How have educational standards been set, by whom, and at what level?

2. What kinds of tests have been used to assess whether the standards are being met?

3. What types of stakes have been attached to tests to ensure that they will be taken seriously?

4. What explicit safeguards have been employed to prevent misuse of test results?

5. Have efforts at raising expectations and checking results brought promise of improved teaching and learning?

Canada's Experience

Question 1

In Canada, educational standards describing what should be taught are set at the province level, with major involvement of educators. There is only a small movement toward national standards (or tests) at this time. Provincial officials clearly want their high-stakes examinations to reflect the present curriculum, and great effort has been expended in employing teachers to help continually redesign the tests so their content stays current. Thus, all five provinces with examination systems have placed a priority on the content validity of the exams at the expense of comparability of exam scores. Each year, new examinations are aligned to the curriculum standards chosen by consensus of selected teachers and provincial agency experts.

The assessments show more variety, but again, all eight provinces use provincial curriculum standards as the starting point for some or all of their assessments. At this time, only one province uses a commercially developed, norm-referenced test that is not linked to the provincial curriculum as its principal assessment instrument. Although provincial education departments give contractors a role in the assessment

programs, most are increasing the role of teachers in determining testing standards, scoring open-ended items, and interpreting results.

Standards and tests, then, are not created apart from provincial and local educators, but by them, along with advice from experts. In general, the involvement of teachers in all phases of standard-setting and testing in Canada is far greater than in the United States. For example, in Alberta, for every 48 students examined, approximately one teacher either writes examination items, serves on a central marking panel, or both; and in Newfoundland, all teachers of biology were surveyed regarding the appropriateness of the specifications for the new biology exam. New York, by contrast—the state that leads all others in terms of the number of tests developed by the state education agency and the number of teachers involved in test development—involves only about one teacher for every thousand students tested (though teachers of Regents courses give feedback afterward via surveys about the examinations).

Question 2

Canadian testing efforts show clearly that there are different tests for different purposes. That is, two entirely different testing systems handle the separate tasks of certifying whether individual students met standards and monitoring whether learning in general, across a province, is in line with what is expected. For the first purpose, all five provinces with examination programs have curriculum-based tests that measure students' achievement in certain courses during their last years of high school. Because the examinations are referenced to the curriculum of particular courses (e.g., biology or world history) rather than covering broad subject areas (e.g., science or social studies), they can provide better coverage of curriculum standards and measure them in more depth.

Of the eight provinces with assessment programs, six use exclusively curriculum-based, criterion-referenced instruments, and two use both normative and criterion-referenced instruments. Only in Newfoundland is a norm-referenced test that is not based upon provincial curriculum used as a principal source of information on provincewide student achievement.

The Canadian provinces have not only established different tests for different purposes, but they have also decided that a wide variety of different kinds of testing methods are to be used, typically combining several in one test. The examinations include a combination of multiple-choice and written-response or open-ended items. Most

provincial assessment programs include multiple-choice, written-response, and other performance components such as portfolios and individual and group participation tasks.

Question 3

At first glance, the five provinces' examination systems appear to show that the stakes for many students are genuinely high. However, the evidence is mixed as to whether these tests really have teeth or not. To be sure, where there are exams, scores partially determine students' final course grades and therefore help determine students' graduation status. Depending upon the province, examination grades contribute between 30 percent and 50 percent of the final course grade. But the contribution of the examination marks is less in provinces with new systems than in provinces that have maintained examination systems since the 1970s. The trend in the new systems is to reduce the examination's contribution and to increase that of teacher grades in hopes of better reflecting the importance of standards that are not measured on examinations. This suggests that one lesson from the Canadian experience is that high stakes are only one way of addressing the goal of improving student learning. Indeed, they may be less necessary when the importance of learning the set curriculum is emphasized in many other ways.

With respect to the assessments, these were designed to measure the overall performance of a group of students in a province. Therefore, no consequences have been attached to the results of these tests for individual students, and most provinces have implemented various safeguards to prevent using assessment results in decisions regarding students' placement and teachers' performance evaluations. Lower stakes ride on these results, but the assessments are regarded as effective in influencing instruction because of their consistency with provincial curricula and wide teacher involvement in their design and implementation.

Question 4

Canadian officials have employed a variety of safeguards to prevent misuse of test results. For the examinations, safeguards are designed to protect students from arbitrary test practices, ensure multiple opportunities for success, and accommodate students with disabilities. Certain procedures built into the examination development systems, such as broad teacher participation and the alignment of the examinations with the curricula, help level the opportunities for students to succeed.

Different sets of safeguards have been developed for the various assessment systems. Safeguards such as aggregated and delayed reporting of scores have been introduced to prevent using scores to certify or place students or to evaluate teachers. Sampling has enabled many provinces to improve the content validity of these tests through the inclusion of performance items to measure skills that cannot be assessed through conventional multiple-choice or written-response items, that may be too expensive to administer to all targeted students, and that may not be accurate for an individual student.

Provincial funding formulas, while not a safeguard per se, have done much to ensure equal opportunity to students in Canada by leveling resources among school districts in a province.

The decision of authorities in most provinces to reinvest annually in test development activities has likely reduced the predictability of test content. These efforts have no doubt increased the ability of the tests to measure students' knowledge of a general domain of content and skills rather than how well they do on specific items familiar from prior tests.

Question 5

For the question regarding demonstrated effectiveness, the answer must be that the jury is still out, since the provinces lack any independent measure showing the effects of the examinations on teaching and learning. On the one hand, provincial testing has been credited with raising standards and improving uniformity in teaching and grading—perhaps necessary steps toward better learning outcomes, but as yet unproven. On the other hand, testing has been criticized for narrowing what is taught and increasing reliance upon teaching approaches closely tied to the tests' specifications. These effects are evaluated differently by different observers.

According to surveys, most students report the examinations are associated with increased motivation to study but also with higher anxiety levels. Thus, the effects on students may be mixed, as some students learn more effectively because of motivation, while others learn less effectively because of anxiety, and still others avoid or minimize possible effects by avoiding demanding examination courses. Finally, some argue that the examinations have increased the fragmentation of the student group by isolating those who take exam courses from those who do not.

And despite the lack of evidence of specific impacts of the sizable current provincial testing in improving educational outcomes, Canada—like the United States—is pursuing additional national testing as well. This effort to supplement the decentralized provincial activities of setting standards and measuring results by establishing a single national test has proven much more difficult and controversial than expected. Despite the difficulties, the mathematics assessment has received final approval from all participating provinces for administration during the 1993 fall semester, and the reading and writing assessment is undergoing final approval for administration during the 1994 spring semester. Some of the sources of controversy in Canada could be expected to be the same in the United States.

Appendix I

Characteristics of Canadian Provincewide Testing Programs

This appendix describes testing programs throughout the Canadian provinces. We have included a summary of the characteristics of testing programs in Alberta and British Columbia, described in chapter 2, to assist readers who may wish to draw comparisons among the 10 provinces. (See tables I.1-I.4.) Descriptions of testing programs elsewhere in Canada are presented to illustrate alternatives to practices in Alberta and British Columbia, although these two provinces are widely regarded as exemplifying practices many other provinces aim to adopt in the future.

Alberta

As noted, chapter 2 discusses Alberta's provincial testing program. Tables I.1 and I.2 summarize characteristics of Alberta's examination and assessment program.

Table I.1: Alberta Diploma Examination System

Category	Characteristic
Stated purposes	To establish high standards of performance
	To certify individual student achievement levels
	To provide a standardized measure of student achievement
	To provide equity in the competition for postsecondary school admissions
Format	Multiple-choice, written-response items
Type	Curriculum-based, criterion-referenced
Students and grades targeted	All students enrolled in examination courses that are usually taken during the senior year of high school
	All English-speaking seniors must write an examination for either an academic or general senior-level English course
Subjects assessed	Biology, chemistry, language arts (academic-level English or general English for English-speaking students and French for French-speaking students), mathematics, physics, and social studies
Scoring	Written-response items are graded by panels of teachers in a central location; multiple-choice items are machine-scored
Safeguards	Iterative review to maximize consensus on validity of content
	Field-testing and analysis of items
	Adjudication processes to allow modifications of examination procedures
	Requirements to accommodate students with disabilities, illness, or bereavement
	Procedures to allow students to initiate a second scoring of an examination
	Opportunities to rewrite examinations with or without retaking the course
Stakes at the student level	Determines 50 percent of student's final grade in examination courses; school-awarded grades determine the other 50 percent
Implementation date	1984

Table I.2: Alberta Assessment Program

Category	Characteristic
Stated purposes	To monitor student achievement on the provincial level
	To provide public accountability
Format	Multiple-choice, written-response, and performance items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	All students in grades 3, 6, and 9 complete the multiple-choice and written-response items; a student sample completes the performance items
Subjects assessed	Language arts, mathematics, science, social studies
	One subject is assessed on a 4-year basis
Scoring	Written-response and performance items are graded by panels of teachers in a central location; multiple-choice items are machine-scored
Safeguards	Spring test results withheld until fall to discourage their use for promotion
	Sampling strategies to increase content coverage
	Field-testing
	Item analysis to reduce gender bias
	Iterative review to maximize consensus on validity of content
Stakes at the student level	None
Implementation date	1982

British Columbia

Chapter 2 describes British Columbia's examination and assessment programs. Future assessments will measure cross-curricular concepts such as communication skills and problem-solving every 2 years. Characteristics of these programs are summarized in tables I.3 and I.4.

Table I.3: British Columbia Provincial Examination System

Category	Characteristic
Stated purposes	To certify students through a standardized measure
	To ensure curriculum is followed
	To respond to public concerns for improved standards
Format	Multiple-choice, written-response items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	All students enrolled in examination courses
Subjects assessed	Biology, chemistry, communications, English, English literature, French (for English-speakers), French (for French-speakers), geography, geology, German, history, Latin, mathematics, physics, Spanish
Scoring	Written-response items are graded by panels of teachers in a central location; multiple-choice items are machine-scored
Safeguards	Iterative review to maximize consensus on validity of content
	Adjudication processes to allow modifications of examination procedures
	Requirements to accommodate students with disabilities, illness, or bereavement
	Procedures to allow students to initiate a second scoring of an examination
	Opportunities to rewrite examinations with or without retaking the course
Stakes at the student level	Determines 40 percent of student's final grade in examination courses; school-awarded grade determines the other 60 percent
Implementation date	1983

Table I.4: British Columbia Provincial Learning Assessment Program

Category	Characteristic
Stated purposes	To monitor achievement at the provincial level
	To provide accountability
	To provide direction for curriculum revision and development
	To direct the allocation of provincial resources
	To direct pre-service and in-service teacher education
	To provide direction for educational research
Format	Multiple-choice, written-response, and performance items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	Samples of students in grades 4, 7, and 10
Subjects assessed	English, mathematics, science, social studies
	One subject is tested every 4 years
	Other subjects (as needed)
Scoring	Written-response and performance items are graded by panels of teachers and contractors in a central location; multiple-choice items are machine-scored
Safeguards	Iterative review to maximize consensus on validity of content
	Sampling strategies to allow adequate coverage of content
	Scores reported on a provincial level only
Stakes at the student level	None
Implementation date	1976

Manitoba

The Manitoba Department of Education reintroduced diploma examinations in 1991.[1] The department plans to follow the model developed by Alberta, but with one major exception: Manitoba will develop and administer examinations in only one subject per year. Although the government initially supported reintroducing examinations in all major subject areas, department officials explained that the costs of providing examinations of acceptable quality in all major academic subjects were too high in comparison to other educational needs. See table I.5 for a summary of this program's characteristics.

[1]Manitoba is a province of about one million people, 60 percent of whom live in Winnipeg, its capital.

Table I.5: Manitoba Provincial Examinations

Category	Characteristic
Stated purposes	To certify student achievement
	To ensure curriculum is followed
Format	Multiple-choice, written-response items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	All grade 12 students enrolled in targeted courses
Subjects assessed	One major subject covered
	Two examinations are developed, one for students in the academic stream and one for students in the general stream
	Each major subject tested every 5 years
Scoring	Written-response items are graded by central panels of teachers; multiple-choice items are machine-scored
Safeguards	Iterative review on validity of content
Stakes at the student level	Determines 30 percent of student's final grade in exam courses
Date implemented	1991

The Manitoba Department of Education has also conducted provincewide assessments in major subject areas at three grade levels on 5- to 6-year cycles. Over the years, different combinations of provincial, school-level, and contracted resources have been used to develop curriculum-based, criterion-referenced tests. Typically, a university professor is hired as a consultant to develop the assessment under the direction of a committee composed of ministry officials and subject-area teachers, with either university students or teachers assisting in developing items and grading the tests. Different sampling plans have been used over the years; currently about 10 percent of students in the subjects and grades targeted for the assessments are tested.

In addition to multiple-choice and written-response items, the Manitoba assessments have included group and individual performance components, student interest questionnaires, and teacher questionnaires and interviews. Various performance components have been used, including individual and group tasks. Student questionnaires collect information about students' interests and attitudes about the provincial curriculum to aid in improving it. Teacher surveys and interviews ask teachers to evaluate the curriculum and gather information about classroom activities and student evaluation problems. Characteristics of this program are summarized in table I.6.

Table I.6: Manitoba Curriculum Assessment Program

Category	Characteristic
Stated purposes	To monitor achievement at the provincial level
	To provide accountability
	To help teachers improve their student evaluation skills
	To improve curriculum and instruction
Format	Multiple-choice, written-response, group performance items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	Samples of students in grades 4, 8, and 10 or 11
Subjects assessed	Language arts, mathematics, science, social studies
	One subject is tested every 5-6 years
	Special assessments in other areas as needed
Scoring	Written-response items are graded by various combinations of teachers and contractors; multiple-choice items are machine-scored
Safeguards	Iterative review to maximize consensus on validity of content
	Results reported on the provincial level only
Stakes at the student level	None
Date implemented	1979

New Brunswick

The provincial authorities established the New Brunswick Achievement Examinations in 1984 as low-stakes tests to monitor student achievement in mathematics and English at grade 11.[2] Under the direction of officials of the ministry of education, two forms of each test were developed—one for students in advanced placement and college preparatory classes and another for students in general or remedial programs. The tests are norm-referenced instruments, which were designed to measure cumulative achievement. About 60-70 percent of the content of the tests is based on the curriculum; the remainder is based on general ability and may not reflect particular curriculum goals. See table I.7 for a description of this program.

Current plans call for the discontinuation of these examinations. New tests are being developed for use during the 1993-94 school year. Unlike the present tests, the new examinations will be criterion-referenced and are expected to contribute 30 percent toward a student's final grade.

[2]New Brunswick is one of Canada's smallest provinces, with a population of about 710,000.

Table I.7: New Brunswick Provincial Achievement Examination

Category	Characteristic
Stated purpose	To monitor achievement
Format	Multiple-choice, written-response
Type	Norm-referenced
Students and grades targeted	Students in grade 11
	One form of the test is provided for students in advanced and college preparation courses; another form is provided for students in general courses
	Tests are optional for students in remedial programs
Subjects assessed	Mathematics and English
Safeguards	Students may request second scoring
Scoring	Centralized scoring of written-response items
Stakes at the student level	Determined at the local level
Date implemented	1984

A criterion-based performance assessment is available for students learning French as a second language.[3] The assessments are voluntary and consist of individual interviews. The New Brunswick educational agency certifies and hires French-language interviewers to administer and grade the interviews, which take about 30 minutes per student, are taped, and often involve two interviewers or one interviewer and one authorized observer. Characteristics of this program are presented in table I.8.

Table I.8: New Brunswick Oral Proficiency Interview for the Second Language Project

Category	Characteristic
Stated purpose	To monitor oral proficiency of French
Format	Oral interview
Type	Criterion-referenced
Students and grades targeted	Students enrolled in French-as-a-second-language courses in grades 3, 6, 9, 10, and 12 on a voluntary basis
Subjects assessed	French oral comprehension (grades 3, 6, 9, 10, and 12)
Safeguards	Certification and training requirements for interviewers
	Interviews are taped to allow rescoring
Scoring	Scored at the school level by trained examiners
Stakes at the student level	None
Date implemented	1978

[3]New Brunswick is second only to Quebec in the size of its French-speaking population—about one-third speak French as a first language.

Newfoundland

Two examination models are in place in Newfoundland.[4] Under the older model, responsibility for writing the provincial examination in a subject is delegated to a teacher. This teacher, using the curriculum and previous examinations as guides, writes the examinations with assistance and supervision from the provincial education department

The Newfoundland Department of Education officially initiated a new approach in 1986 with the development of a new biology examination with comprehensive changes intended to improve its quality. The department involved large numbers of teachers to develop consensus on the specifications and to write items.[5] A university-based consultant directed the development of a large pool of examination questions at diverse levels of difficulty that could be used over a period of years to develop comparable examinations and avoid the expense of developing new items each year and the need to adjust for differences in the difficulty of examinations from year to year.

Newfoundland officials shrank the testing program from 33 examinations to 18 in order to devote more resources to test revision and to reduce the logistical problems resulting from administering a large number of examinations. Even so, funds are inadequate to support similar major revision efforts. The new approach no doubt produced better examinations; however, the costs associated with using a large number of participants and with developing a large item bank were not feasible. Characteristics of the Newfoundland and Labrador examinations are presented in table I.9.

[4]The province of Newfoundland, which includes Labrador, is a large but sparsely populated area with only 550,000 inhabitants.

[5]All biology teachers in the province played some role in the development of the new biology examination.

Table I.9: Newfoundland and Labrador Public Examinations

Category	Characteristic
Stated purposes	To certify student through standardized reporting of curriculum
	To provide postsecondary schools with reliable information for making admissions decisions
	To provide public accountability
Format	Multiple-choice, written-response items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	Students enrolled in selected academic courses
Subjects assessed	Biology, business English, chemistry, English (2 levels), environmental sciences, French, geology, mathematics (3 levels), physics, thematic literature (3 levels), world geography, world history, world problems
Scoring	Written-response items are graded by panels of teachers in central locations; multiple-choice items are machine-scored
Safeguards	Procedures to allow students to request a second scoring
	Opportunities to rewrite examinations with or without retaking the course
Stakes at the student level	Exam grades determine 50 percent of students' grades in examination courses
	Exam moderates school-awarded grade (school-awarded grades are adjusted according to the degree of agreement between the average examination score and the average school-awarded grade for each class grouping)
Date implemented	1974

The Newfoundland Department of Education administers commercially developed, norm-referenced assessments to determine student achievement in three core subjects at various grade levels. The test includes only multiple-choice items and is completely scored by machine. (See table 1.10 for details.) The department developed grade 6 science and grade 3 mathematics tests that are linked to the provincial curriculum for use along with its norm-referenced program.

Table I.10: Newfoundland and Labrador Achievement Testing Program

Category	Characteristic
Stated purposes	To monitor achievement To provide public accountability
Format	Multiple-choice
Type	Norm-referenced
Students and grades targeted	Students in grades 4, 6, and 8 every 3 years and a sample of students in grade 12 annually
Subjects assessed	Language arts, mathematics, and social studies
Scoring	Machine-scored
Safeguards	None
Stakes at the student level	Determined at the local level
Implementation date	1974

Nova Scotia

In 1972, the program of provincial examinations was replaced with the Nova Scotia Achievement Tests.[6] Each year, all students in grades 9 and 12 take a multiple-choice, norm-referenced test that measures skills in seven areas. The tests are developed by the provincial education department and are based on the provincial curriculum. Each year, 25 percent of the items on the test are rewritten by subject-area teachers. (See table I.11 for details concerning these tests.) Beginning in the 1988-89 school year, the province developed a criterion-referenced, written-response mathematics assessment for students in grade 5. A language arts assessment, consisting of a review of students' writing spanning grade 5, was being developed at the time of our review. Provincial officials, consultants, and subject-area teachers play key roles in the development of the grade 5 tests.

[6]Nova Scotia is a compact province with a population of about 900,000.

Table I.11: Nova Scotia Achievement Tests

Category	Characteristic
Stated purposes	To monitor achievement
	To provide public accountability
Format	Multiple-choice items
Type	Norm-referenced
Students and grades targeted	Students in grades 9 and 12
Subjects assessed	Seven broad areas: language usage, mathematics applications, mathematics essentials, proofreading, reading, science, social studies
Scoring	Machine-scored by an external agency
	Analysis and reports prepared by an external agency
Safeguards	25 percent of items are rewritten each year
Stakes at the student level	None
Date implemented	1972

Ontario

There is no central examination program in Ontario.[7] However, districts must develop examinations that measure the achievement of high school students in specified advanced high school courses. The provincial ministry of education specifies the curriculum for these courses, provides workshops in curriculum and assessment for teachers, and monitors the grading of district examinations by independently regrading a sample of examinations submitted by each school in each subject.

On the assessment side, Ontario Ministry of Education introduced a Provincial Review Program in 1986. The ministry determines the subjects that will be assessed and the grades targeted for the assessment. Test development work is typically contracted out to a local school district. The tests are criterion-referenced instruments with both open-ended and multiple-choice items. Performance items requiring teacher observations were introduced in the 1989 mathematics assessment, and centralized scoring was introduced in 1992. See table I.12 for a description of these tests.

[7]Ontario is the largest province in Canada, with a population of 9 million. In general, the school districts in Ontario have more specialists in areas such as curriculum and testing than do districts elsewhere in Canada. This has lessened the need for provincial testing services.

Table I.12: Ontario Student Assessments

Category	Characteristic
Stated purposes	To provide public accountability
	To improve curriculum
Format	Multiple-choice, written-response, performance items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	Sampling is used for students at different grades in major subjects according to need and the availability of funds
	At the elementary level, a representative sample of students from 100 English-language schools and 100 French-language schools are selected to participate; at the high school level, all schools are included
	Districts may choose to be included for a small fee (about $175) and have all students enrolled in a targeted course of studies tested
Subjects assessed	As determined by the ministry
Scoring	Beginning in 1992, written-response items scored by central grading panels
Safeguards	Grades are not reported for individual students
Stakes at the student level	None
Implementation date	1986

Prince Edward Island

Prince Edward Island does not now administer any provincial tests.[8] After the examination system was discontinued in the 1970s, the provincial education department administered a commercial, norm-referenced test to students at various grade levels. This program was discontinued in 1992. Officials told us the tests were not useful, that their costs were not matched by the value of information gained, and that they could be superseded if the province uses the Canada-wide assessments being developed by the Council of Ministers of Education, Canada (as discussed in chapter 5).

Quebec

Like Newfoundland, Quebec has had a long-standing, uninterrupted examination program.[9] The Quebec Ministry of Education introduced student assessment for the purpose of program evaluation in 1979.

[8]Prince Edward Island has the smallest area and, with about 130,000 residents, the smallest population of the provinces.

[9]Quebec is the second most populous province in Canada with a population of about six and one-half million.

Over the past years, the ministry has increasingly involved teachers in examination development. Currently, between 500 and 700 teachers, under the direction of the ministry of education, write, revise, and translate examinations. As elsewhere, test items are developed to reflect content emphasized in the provincial curricula. Again, for test security, items are not field-tested; panels of teachers and ministry officials validate items through discussion and consensus. The ministry of education approves the examinations without the involvement of external committees or a governing board. In contrast to practices in provinces with newer examination systems (Alberta, British Columbia, and Manitoba), open-ended examination items are graded at each school by classroom teachers using a set of guidelines developed by the provincial education ministry. The ministry regrades a small sample of examinations to check accuracy and consistency and can direct regrading of all tests from a school.

Fewer examinations are now administered in Quebec, as the ministry dropped exams for vocational education courses to concentrate on improving academic course exams. From 104 examinations given in 1988, the number dropped to 21 in 1990 (12 for schools where French is the language of instruction and 9 for schools teaching in English). See table I.13 for a description of these examinations.

Table I.13: Quebec Ministry-Prepared Examinations

Category	Characteristic
Stated purposes	To certify students on the basis of a standardized measure
	To provide postsecondary schools with reliable information for making admissions decisions
	To provide public accountability
Format	Multiple-choice, written-response items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	Students enrolled in examination courses
Subjects assessed (in French)	Academic chemistry, advanced chemistry, academic physics, advanced physics, academic mathematics, advanced mathematics, history, geography, economics, French composition, French literature, oral French
Subjects assessed (in English)	Academic chemistry, advanced chemistry, advanced physics, academic mathematics, advanced mathematics, history, geography, economics, English language arts
Scoring	Schoolhouse scoring of written-response items
Safeguards	Items are analyzed to determine appropriateness
	Scores are adjusted to control for differences in pass rates from one year to the next
	The ministry rescores a sample of examinations from some schools to monitor scoring consistency
Stakes at the student level	Determines 50 percent of final grades in courses where exams are required; the moderated school-awarded grade contributes the other 50 percent
	Exam moderates school-awarded grades (school-awarded grades are adjusted according to the degree of agreement between the average examination score and the average school-awarded grade for each class grouping)
Implementation date	1929

The province does not conduct regular provincial student assessments. Rather, student assessment typically is conducted to coincide with the introduction of a new curriculum or gather information about achievement in subjects to address the need to revise the course of studies. The provincial education agency determines which subjects will be assessed and specifies which curriculum priorities will be tested. Typically, tests include written-response and multiple-choice questions and are administered to representative samples of students. Centralized scoring was used for the first time to grade the 1991 writing assessment, a popular

change that will likely be repeated in the future. See table I.14 for a description of the assessments.

Table I.14: Quebec Curriculum Tests

Category	Characteristic
Stated purposes	To evaluate effectiveness of the curriculum
	To provide public accountability
Format	Multiple-choice, written-response items
Type	Criterion-referenced, curriculum-based
Students and grades targeted	No set cycle
Subjects assessed	As needed
Scoring	Most recent assessment (writing, 1991) scored by central panels of teachers
Safeguards	Scores are not reported on the individual level
Stakes at the student level	None
Implementation date	1979

Saskatchewan

Saskatchewan has neither a provincial examination system nor an assessment program in place.[10] A new curriculum-based, criterion-referenced assessment system is planned for implementation in 1993 to coincide with the introduction of a new science curriculum. Additional assessments will be developed as new curricula are introduced in other subject areas. Assessment development will be largely in-house, with teachers and administrators participating in specification and item development, administration, scoring, and interpretation of results.

[10]Saskatchewan is a mid-sized province with a population of about 1 million.

Expert Participants in Site and Telephone Interviews

We wish to acknowledge the cooperation and assistance in gathering information that was provided to us by the following individuals.

John O. Anderson, Professor, Department of Education, University of Victoria, British Columbia

Leonard Babcock, Director, Student Assessment Branch, Department of Education, St. John's, Newfoundland

David Bateson, Professor, Department of Education, University of British Columbia, Vancouver

Monique Belanger, Coordinator, Council of Ministers of Education, Canada, Toronto

Dennis Belik, Coordinator of Achievement Assessments, Student Evaluation Branch, Alberta Education, Edmonton

David Bond, Vice President of Public Affairs, Hong Kong Bank of Canada, Vancouver

Peter Calder, Professor, Department of Educational Psychology, University of Alberta, Edmonton

Robert K. Crocker, Professor, Institute for Educational Research and Development, Memorial University of Newfoundland, St. John's

Jim Cullen, Director, Council of Ministers of Education, Canada, Toronto

Thomas Dunn, Coordinator of Diploma Examinations in the Humanities, Student Evaluation Branch, Alberta Education, Edmonton

John Eastaugh, Coordinator of Diploma Examination, British Columbia Department of Education, Victoria

Keith Gray, Vice President of Government Relations and Educational Services, Business Council of British Columbia, Vancouver

Cary Grobe, Director of Student Evaluation Branch, New Brunswick Department of Education, Fredericton

Nadia Hochachka, Assistant to the Director of the Student Evaluation Branch, Alberta Education, Edmonton

Frank G. Horvath, Director of the Student Evaluation Branch, Alberta Education, Edmonton

Ivan Johnson, President of the British Columbia Association of Teachers of Mathematics, Vancouver

Yvonne Johnson, Coordinator of Mathematics and Science Diploma Examination Program, Student Evaluation Branch, Alberta Education, Edmonton

Michael Kozlow, Education Officer, Program Implementation and Review Branch, Ministry of Education, Toronto

Leo Laroche, Consultant to Evaluation Unit, Quebec Ministry of Education

Ray Malone, Pupil Personnel Consultant, Prince Edward Island Department of Education, Charlottetown

Norman Mayer, Director of Curriculum Development and Implementation, Manitoba Education, Winnepeg

Becky Matthews, Acting Director of Assessment, Examinations, and Reporting Branch, British Columbia Ministry of Education, Victoria

Turney Manzer, Assistant Director of Research, Nova Scotia Department of Education, Halifax

David McCamus, Chairman of Xerox Canada, Ltd., North York, Ontario

Thomas O'Shea, Professor, Department of Education, Simon Fraser University, Burnaby, British Columbia

Todd Rogers, Professor, Department of Educational Psychology, University of Alberta, Edmonton

Martin Ruane, Evaluation Consultant to Evaluation and Student Records, Saskatchewan Education, Regina

Elana Scraba, Assistant Director of Student Evaluation Branch, Alberta Education, Edmonton

Bill Toth, Coordinator of Provincial Learning Assessment, Assessment, Examinations, and Reporting Branch, British Columbia Ministry of Education, Victoria

Ross E. Traub, Professor, Ontario Institute for Studies in Education, Toronto

Paul Vachon, Director of Evaluation Unit, Quebec Ministry of Education

Daina Watson, Assistant Director of Assessment, Examinations, and Reporting Branch, British Columbia Ministry of Education, Victoria

Marvin F. Wideen, Professor, Department of Education, Simon Fraser University, Burnaby, British Columbia

Major Contributors to This Report

Program Evaluation and Methodology Division

Frederick V. Mulhauser, Assistant Director
Kathleen D. White, Project Manager
Christine Ing, Research Associate
Venkareddy Chennareddy, Referencer

Bibliography

Anderson, John O., et al. The Impact of Provincial Examinations on Education in British Columbia: General Report, ERIC Document ED 325516. Victoria: British Columbia Department of Education, 1990.

Calder, Peter. Impact of Diploma Examinations on the Teaching-Learning Process. Edmonton: Alberta Teachers' Association, 1990.

Crocker, Robert. Public Examinations Item Banking Project. St. John's: Newfoundland and Labrador Department of Education, 1989.

Government of Canada. Education in Canada. External Affairs and International Trade, Canada, 1989.

Koretz, Daniel M., et al. The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education. Chicago: April 5, 1991.

National Council on Education Standards and Testing. Raising Standards for American Education. Washington, D.C.: 1992.

U.S. Congress, Office of Technology Assessment. "How Other Countries Test." Chapter 5 in Testing in American Schools: Asking the Right Questions, OTA-SET-519. Washington, D.C.: U.S. Government Printing Office, February 1992.

Wideen, Marvin F., et al. Impact of Large-scale Testing on the Instructional Activity of Science Teachers. Burnaby, British Columbia: Institute for Studies in Teacher Education, Simon Fraser University, 1991.

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1000
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066.

PRINTED ON RECYCLED PAPER

BEST COPY AVAILABLE